DOCUMENT RESUME

ED 195 576 TM 810 026

TITLE Writing Assessment for the 1980s. Proceedings of a National Conference on the Assessment of Writing (Boulder, Colorado, June 13, 1980).
INSTITUTION Northwest Regional Education Lab., Portland, Oreg. Clearinghouse for Applied Performance Testing.
SPONS AGENCY National Inst. of Education (DHEW), Washington, D.C.
PUB DATE 13 Jun 80
NOTE 110p.

EDRS PRICE MF01/PC05 Plus Postage.
DESCRIPTORS Elementary Secondary Education: *Evaluation Methods: Higher Education: *Scoring: Teacher Workshops: *Writing (Composition): *Writing Skills
IDENTIFIERS Analytic Scoring: Generalizability Theory: Holistic Evaluation: National Assessment of Educational Progress: Primary Trait Scoring

ABSTRACT

The proceedings of a one-day training seminar on the assessment of writing proficiency are presented. The seminar was designed to bring educators up to date on newly developed assessment methods and to help them see the developmental directions of the future. The program was divided into two parts: large-group presentations and small-group training sessions. Texts of the three presentations were reproduced in their entirety and dealt with: (1) a challenge to measurement specialists to make writing assessments relevant and useful to the classroom teacher: (2) an analysis of the similarities and differences between direct (writing samples) and indirect (objective tests) measures of writing: and (3) some possibilities for future approaches to writing assessment. Each of the four training sessions offered a practical set of hands-on experiences to help participants conduct better writing assessment. Session one provided participants with instruction and practice in conducting holistic, analytical, and primary trait scoring. Session two provided holistic scoring in greater depth. Session three provided primary trait scoring in greater depth. Session four provided participants with instruction in the application of generalizability theory to the reliability of writing sample ratings. (RI)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED195576

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

WRITING ASSESSMENT FOR THE 80s

Proceedings of a Training Seminar presented by the Clearinghouse for Applied Performance Testing

Northwest Regional Educational Laboratory

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

M. Thorne

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

TM 810 026

WRITING ASSESSMENT FOR THE 1980s

Proceedings of a National Conference
on the Assessment of Writing

presented by

Clearinghouse for Applied Performance Testing

in collaboration with

National Assessment of Educational Progress

Boulder, Colorado
June 13, 1980

Prepared by:
Clearinghouse for Applied Performance Testing
Northwest Regional Educational Laboratory
710 S.W. Second Avenue
Portland, Oregon 97204

This project is supported through a grant
from the National Institute of Education,
Department of Education. However, the
opinions expressed herein do not necessarily
reflect the positions or opinions of NIE
and no official endorsement by NIE should be inferred.

WRITING ASSESSMENT FOR THE 1980s

Seminar Instructional Staff

J. Evans Alloway
Director, Programs for the Assessment of Writing
Educational Testing Service
Princeton, New Jersey 08541

Beverly Anderson, Ph.D.
Director, Assessment and Measurement Program
Northwest Regional Educational Laboratory
710 S.W. Second Avenue
Portland, Oregon 97204

Rexford Brown, Ph.D.
Director, Publications and User Projects
National Assessment of Educational Progress
1860 Lincoln Tower
Suite 300
Denver, Colorado 80295

Robert Brennan, Ed.D.
Director, Department of Measurement Research
The American College Testing Program
2201 North Dodge Street
P.O. Box 168
Iowa City, Iowa 52243

Edmund Farrell, Ph.D.
Professor of English Education
College of Education
University of Texas at Austin
Austin, Texas 78712

Ina Mullis, Ph.D.
Senior Research Analyst
Analysis and Research
National Assessment of Educational Progress
1860 Lincoln Tower
Suite 300
Denver, Colorado 80295

Richard Stiggins, Ph.D.
Coordinator, Clearinghouse for Applied Performance Testing
Northwest Regional Educational Laboratory
710 S.W. Second Avenue
Portland, Oregon 97204

ACKNOWLEDGEMENT

Special recognition is due to Rick Stiggins who was responsible for planning, arranging and coordinating the seminar. CAPT would also like to offer thanks to Jack Schmidt and Barbara Lohne of the NAEP staff for their valuable assistance in the planning and organization of this training seminar. Recognition is due to Nancy Bridgeford and Carol DeWitte of the CAPT staff for their important contributions to the seminar effort. Further, CAPT would like to express special appreciation to the seminar instructors for their thorough preparation and excellent presentation of practical and useful training information. And, finally, thanks are due to the seminar participants for their contribution to a successful and stimulating seminar.

Beverly Anderson
CAPT Director

Copyright © 1980 by the
Northwest Regional Educational
Laboratory. All rights reserved.

TABLE OF CONTENTS

	Page
INTRODUCTION	1
Background Information	1
About the Proceedings	3
PART ONE: LARGE GROUP PRESENTATIONS	5
Writing Assessment - The Teacher's Perspective Edmund Farrell University of Texas	6
Alternative Methods of Writing Assessment Richard Stiggins Clearinghouse for Applied Performance Testing	20
New Frontiers in Writing Assessment Rexford Brown National Assessment of Educational Progress	47
PART TWO: SMALL GROUP TRAINING SESSIONS	62
Introduction to Section	63
An Introduction to the Scoring of Writing Samples Beverly Anderson Richard Stiggins Clearinghouse for Applied Performance Testing	64
Holistic Evaluation of Students' Writing J. Evans Alloway Educational Testing Service	73
Primary Trait Scoring Ina Mullis National Assessment of Educational Progress	79
Application of Generalizability Theory to Examining the Dependability of Writing Sample Ratings Robert Brennan American College Testing Program	90
APPENDIX	
List of Seminar Participants	94

INTRODUCTION

Background Information

On June 13, 1980, the Clearinghouse for Applied Performance Testing (CAPT) presented a one-day training seminar on the assessment of writing proficiency. The seminar was presented in collaboration with the National Assessment of Educational Progress and was offered as a post-session to NAEP's Tenth Annual Conference on Large-Scale Assessment in Boulder, Colorado. The training session brought together prominent specialists in writing assessment with over 90 educators and measurement specialists from across the country for an indepth discussion of current and future issues in writing assessment.

Because of a growing emphasis on basic skills instruction, including writing over the past decade, and far-reaching concern regarding an apparent nationwide decline in writing skills, educators have focused major attention on writing instruction. Special programs designed to teach writing skills have emerged at all educational levels. Since assessment is an integral part of the instructional process, issues of assessment had to be addressed in the context of each of these new programs.

Due to the serious shortage of assessment methods early in the decade, major developmental efforts were launched to devise new methods. As a result of those efforts, major strides have been made in developing and implementing new approaches. The CAPT seminar, Writing Assessment in the 1980s, was designed to bring educators up to date on newly developed assessment methods and to help them see the developmental directions of the future.

More specifically, the conference was intended to achieve two objectives. The first was to provide participants with information on

current developments in writing assessment methodology. The second was to offer a series of very practical hands-on training experiences in methods of writing assessment. The program planned to achieve these objectives was divided into two parts: large-group presentations in the morning and four small-group training sessions in the afternoon. The conference program was structured as follows:

AM	LARGE GROUP PRESENTATIONS:
8:30 - 9:00	Welcoming Remarks, Introductions, and Overview of the Seminar Beverly Anderson, Clearinghouse for Applied Performance Testing
9:00 - 10:00	WRITING ASSESSMENT - The Teacher's Perspective Edmund Farrell, University of Texas
10:15 - 11:00	Alternative Methods of Writing Assessment Richard Stiggins, Clearinghouse for Applied Performance Testing
11:00 - 12:00	New Frontiers in Writing Assessment Rexford Brown, National Assessment of Educational Progress

PM	TRAINING SESSIONS:
1:15 - 2:45	Session #1: An Introduction to the Use of Writing Samples Beverly Anderson and Richard Stiggins Clearinghouse for Applied Performance Testing
	Session #2: Holistic Evaluation of Students' Writing Evans Alloway, Educational Testing Service
	Session #3: Primary Trait Scoring Ina Mullis, National Assessment of Educational Progress
	Session #4: Applications of Generalizability Theory to Examining the Dependability of Writing Sample Robert L. Brennan, American College Testing Program
3:00 - 4:30	Four training sessions repeated a second time

About the Proceedings

The proceedings of the seminar are summarized here. The text of the three morning presentations are reproduced in their entirety, and abstracts of the four training sessions are provided, along with samples of the training materials.

In the first paper from the morning session, Professor Farrell, of the University of Texas, issues a strong challenge to the measurement community to make writing assessments more relevant and useful to the classroom teacher in planning and conducting writing instruction. Looking to the future, he calls for the abolition of large-scale writing assessments if they fail to serve teachers' needs.

In the second presentation, Dr. Stiggins, of the CAPT staff, provides a detailed analysis of the similarities and differences between direct (via writing samples) and indirect (via objective tests) measures of writing. Both approaches are considered by the author to be viable and useful measurement strategies, and those concerned with writing assessment are encouraged to take advantage of the strengths of each assessment approach.

In the third paper, Dr. Brown, of NAEP, speculates on the future direction of writing assessment. Acknowledging that numerous gaps in our knowledge regarding writing assessment still exist, Brown argues that in the future we will need to probe further into the cultural context of writing, into the changing attitudes toward writing, and the component skills and developmental patterns of writing. The author also describes some possibilities for future approaches to writing assessment.

Each of the four training sessions presented in the afternoon offered a practical set of hands-on experiences to help the participants

conduct better writing assessment. Sessions were repeated twice, allowing participants to attend two of the four sessions.

In the first training session, Beverly Anderson and Richard Stiggins, of CAPT, presented an overview of a few of the most frequently used strategies for evaluating writing samples. Participants received instruction and practice in conducting holistic, analytical and primary trait scoring. Training session two, presented by Evans Alloway, of ETS, provided more detailed instruction on and practice with holistic scoring. In training session three, Ina Mullis, of NAEP, offered practical instruction on the development and use of primary trait scoring. And in the fourth training session, Bob Brennan, of ACT, instructed participants in a psychometric issue regarding the application of generalizability theory to examine the dependability (reliability) of writing sample ratings.

PART ONE: PRESENTATIONS

- Writing Assessment - The Teacher's Perspective
- Alternative Methods of Writing Assessment
- New Frontiers in Writing Assessment

WRITING ASSESSMENT - THE TEACHER'S PERSPECTIVE

Edmund J. Farrell
University of Texas, Austin

If classroom teachers of English are not enamored of present assessments of writing--and both my reading of the professional literature and my consultancies with the schools inform me that they are not--their dissatisfaction is understandable. The assessments are inadequate to the task they presume to undertake: the tests they employ place far too much emphasis on students' abilities to edit the writing of others rather than on their ability to compose their own thoughts and feelings in writing; further, on those rare instances when a test does call for actual pieces of writing, it calls for far too few over far too short a period of time, rendering highly suspect any subsequent interpretation of students' ability to write.

Essential Writing Skills

In February 1976, along with seven other specialists in the English Language Arts, I had the privilege of participating in a two-day invitational conference called by the American College Testing Program

(ACT) to review the conceptual structure of the language arts curriculum and, indirectly, to advise ACT on its language assessment programs. Though the eight consultants represented various fields of specialization and institutional affiliation, we were of one mind in giving a low priority in the curriculum to tasks of editing, not that such tasks are unimportant but that they are not nearly so important as helping students both to find something worth saying and to say it in writing that has form, fluency, and style.

In her report of that invitational conference, What's Really Basic in the Language Arts?, Renee M. Huntley writes:

> Given the necessity to emphasize in the classroom the content matter of how language works--socially, intellectually, psychologically--and in terms of that major mission, usage/mechanics was invariably conceded to be relatively superficial. However, since the participants were asked to weigh their responses in terms of the overall usage/mechanics skills deemed minimally acceptable in college, they recognized that colleges do in fact reward the ability to process the printed word in the standard language (often referred to as Edited American English). But, they maintained the testing of usage is the testing of social class, not of educational achievement. It implies that success in college is not based on knowledge or "language-thinking," but on social background...[Edited American English] is not in itself an absolute, unchanging standard but a dialect, a convention whose lack should not signal inferiority. (Huntley, p. 6)

Evidence that college departments of English, while recognizing the importance of mechanics and usage, do not place high priority on these skills was furnished by studies conducted by Educational Testing Service (ETS) to ascertain whether items on the Test of Standard Written English (TSWE) reflected concerns shared by the English profession (A Preliminary Study of the Test of Standard Written English, Breland Hunter, et al, ETS, 1976). ETS researchers in 1973 surveyed a sample number of English departments in two-year and four-year colleges in an attempt to find out what was being taught and what kinds of tests departments would find most useful. To supplement the survey, questionnaires containing forty items

were sent to 200 professors of English in 200 separate colleges. Professors were asked to indicate whether each of the items, which ranged from "using transition words correctly" to "expressing the writer's own feelings," was "very important," "of some importance," "not important," or "irrelevant" to instruction in freshman English. The researchers also examined journals such as the English Journal, College English, PMLA, and Research in the Teaching of English to determine whether contributors placed great importance on writing skills being measured by TSWE. The answers from all three sources were essentially the same: though skills measured by TSWE were desirable, they were not the most important considerations in teaching or assessing writing.

The top five priorities of English professors were these: (1) writing a unified essay; (2) using supporting detail; (3) arranging arguments logically; (4) making verbs agree with subjects; (5) writing expository prose. Of those five, TSWE treats only "making verbs agree with subjects." Nevertheless, items on TSWE are not devoid of importance: eleven of the skills measured by the test rank among the seventh to the twentieth most important concerns cited by English professors.

Typical of current machine-scored tests that place undue emphasis on tasks of editing is that titled "A Writer's Skills," developed by ETS for its Basic Skills Assessment program. The test has 14 items on spelling, 14 on capitalization and punctuation, 4 on filling out forms correctly, 16 on usage, and 27 on evaluation and organization of writing. In short, of 75 items, 48 have to do with editing skills. Although the primary purpose of tests developed for the Basic Skills Assessment is ostensibly to identify students requiring remediation in basic skills, one could argue that skills being assessed are not truly basic and that, even if they were, the tests yield the wrong kind of information to assist a classroom

teacher in helping those who write poorly. The items place a premium on textbook-type deductive exercises rather than on practice in writing, proficiency in which will never be attained unless one writes.

In 1975 the National Assessment of Educational Progress published Writing Mechanics, 1969-1974: A Capsule Description of Change in Writing Mechanics as revealed in the papers of 9-, 13-, and 17-year olds who had participated in either the 1969 or 1974 assessment. Richard Lloyd-Jones, chairman of the English Department, University of Iowa, and Ross Winterowd, professor rhetoric, University of Southern California, critiqued the report and made a number of observations about its implications, among them, these:

> We have to recognize that teaching grammar is not teaching writing. Grammar is an important subject in its own right; there is no connection, however, between ability to describe language and ability to use it.
>
> If we want better writing, we need to require more of it; if we require more of it, we need more full- or part-time people to respond constructively to what is being written. (p. 45)

The call for more individuals to respond to students' writing stems from the authors' knowledge that a secondary teacher of English with 150 students faces 25 hours of reading compositions each time he or she elects to spend 10 minutes per student evaluating papers. When one considers that most teachers spend 25 hours a week in the classroom; make two to three preparations per day for classes; have departmental meetings, faculty meetings, PTA meetings to attend; and often supervise extracurricular activities, the burden of responding frequently and carefully to students' writing can be viewed as truly awesome.

With regard to "the basics," Lloyd-Jones and Winterowd comment, "There is no evidence here that the schools must 'go back to the basics'; indeed, the basics seem to be well in hand" (p. 43), a judgment forshadowed in

Chapter 1 of the report, "Abstracts of Results." There one reads the following about the papers of 17-year olds:

> In general, most of those aspects of writing generally called "mechanics" and stressed heavily in elementary and junior high school English classes (e.g., punctuation, capitalization, agreement, spelling, word usage and so on) are being handled adequately by the vast majority of students, and there is no evidence of deterioration in their use. (p. 1)

A Strong Tie Between Writing and Reading Skill

Rather than continue to place strong emphasis on tests of mechanics and usage, those interested in assessing students' writing abilities might get closer to the heart of the matter through tests of critical reading. Though no substitute for writing samples, the tests seem to complement them in subtle ways, a fact realized by those who have yearly authored the Advanced Placement Examination in English.

About the relationship between reading and writing, participants to ACT's 1976 Invitational Language Arts Conference

> ...agreed that writing samples would clearly be the best means for assessing writing ability. However, they did acknowledge that writing and critical reading were inextricably related. Though the exact nature of that relationship is yet to be determined through research, the participants felt that exercises in critical reading would be a suitable passive way to assess writing potential. They did, however, caution against an equal weighting of the components of critical reading in any such assessment. For example, there is a category of inferential understanding so that more items should be devoted to "recognizing the author's purpose and motives" and "identifying the author's underlying assumptions."...than to other skills in that category. (What's Really Basic in the Language Arts, p. 16)

In summarizing what results of the 1969 and 1974 NAEP assessments of writing might imply, Lloyd-Jones and Winterowd comment, "Writing is inextricably tied to reading: good readers are very often good writers and vice versa. If 1974's average 13- or 17-year-old has done less reading than 1969's, he or she could be expected to be a somewhat poorer writer." (Writing Mechanics, 1969-1974, p. 43)

In Teaching Expository Writing, William Irmscher credits E.D. Hirsch, Jr., director of freshman English at the University of Virginia, with having made a highly succinct observation about the connection between reading and writing: "In a speech made at a writing conference in 1977, E.D. Hirsch, Jr., said (I quote him from my notes), 'You cannot possibly write better than you read.'" (p. 157)

Deficiencies in Direct Writing Assessments

As I earlier indicated, assessments that do elicit writing from students usually suffer from a number of deficiencies: topics are often poorly designed; students are given too little time in which to compose, and too few pieces of writing are called for. Unfortunately, the public--including legislators and parents--are frequently led to believe that deficiencies in assessment are deficiencies in instruction.

Contemporary rhetoricians agree that a central task for any writer is that of finding an appropriate narrative or speaking voice, a voice that will be shaped by the audience, the purpose, and the occasion for the writing. The Gettysburg Address is what it is--an inspired American oration--because Lincoln's choice of language was largely determined by his knowledge of where, why, and to whom he would be speaking. Yet many students whose writing is being assessed for purposes of placement, as with the ETS English Composition Test, or purposes of commencement, as with some state assessment programs, are asked to address themselves to unknown audiences for undefined purposes under artificial circumstances or occasions. Furthermore, they are expected to produce a coherent piece of written discourse under the inexorable tyranny of the clock--usually in twenty minutes, less time than it takes most professional writers to warm up to a subject. In short, the topics as well as the conditions under

which students are expected to respond to them defy what teachers of English say about writing as a recursive process--one that involves discovering, arranging, and revising one's thoughts in the light of particular occasions, audiences, and purposes.

Problems With NAEP's Writing Assessment

The artificiality of NAEP writing assignments, which are better than most, has been criticized by both John Mellon, professor of English at the University of Illinois at Chicago Circle, and Paul Olson, professor of English at University of Nebraska. In National Assessment and the Teaching of English, Mellon writes of the meaninglessness of filling out a job application when no job is in the offing or of doing exercises in which one is "asked to write by examiners whom [one] does not know, told that [one's] teachers will not see [one's] writing, that it would not influence one's marks or academic futures, and [one] will receive no feedback at all on [one's] efforts." (p. 34) In a pamphlet published under the aegis of the North Dakota Study Group on Evaluation and titled A View of Power: Four Essays on the National Assessment of Educational Progress, Olson reprints the following NAEP essay topic assigned to 9-year olds, one illustrated by a picture depicting a forest fire with animals swimming across the river rapids to safety:

> Here is a picture of something sad that is going on in the forest. Look at the picture for a while. Do you see the forest fire? Write a story about what is happening in the picture. This is an important story because you want people to know about this sad event. (Emphasis Olson's)

Olson then offers the following analysis of the assignment:

> First of all, notice that the directions do not place the spectator and writer in a situation which approaches similitude. No one seeing a forest go up in smoke would go write a story--certainly not a child. "Help" is the cry. And no one seeing a picture of a forest fire would write "An important story because (he or she) wants people to know about this sad event." Children--at least the children I

> know--are likely to ask how the fire got started: "Did the animals get out"? What do I do if I get caught like that? Will I die?" etc. Given the assignment's language (story? event?), it is not clear whether the child is being asked to pretend that the picture is a fictive picture which he is to frame with a story as in a children's book, or a real picture demanding that he respond to it--an important test of his capacity to use language in crises, to produce action. (p. 17)

Later in the pamphlet, Olson presents his analysis of a NAEP writing assignment intended for 13-year olds, that of inviting the mayor of one's city to speak at one's school:

> Placing before students unsituated tasks leads to unsituated responses and some really silly data...The invitation is meant to tell the mayor where the school is, suggest his topic, and set the date when he is to speak. The National Assessment indicates that such an answer as "I invite you to speak about your job to my class at Bryan Jr. High School whenever you can come," would be an acceptable answer. That may be an acceptable answer on a written test, but I know of no high school that would permit such an invitation to go out. Moreover, few 13-year old students are confronted with having to invite a mayor to speak at their schools, with suggesting his topic, and with discussing the choice of speaking dates with him. According to the assessment, only 27 per cent of the students could do that task. Why? Because they were inadequate to the task or because it is a nontask? (p. 20)

Teachers of English who have been left with the responsibility of interpreting NAEP findings for whatever curricular implications they yield, have been perplexed by just such a question--if the writing appears to be faulty, to what extent is the cause in the respondent and to what extent in the stimulus?

However they answer that query, teachers who read What Students Know and Can Do: Profiles of Three Age Groups (NAEP, 1977) will be left wondering what to do on Monday morning to make competent scribes of their charges. What are they to make of such tidbits about 9-year olds as the following: "Many can write without making punctuation, agreement or word-choice errors." "Many are familiar with the basic conventions of letter writing and can write letters to friends." "Some are able to write thank-you notes." "Some can compose essays when given explicit

instructions." "Few can organize and elaborate ideas in writing." "Few can revise their own writing to improve its overall organization." "Few can write essays free of misspelled words, sentence fragments and run-on sentences." Even if they could trust the findings, even if the findings were not couched in such vague generalizations, teachers would be left to resolve for themselves the distinctions between knowing what students can do and knowing what they should do to become better writers. Any teachers worth their salt already know which kids in class cannot write well; what they do not always know is what to do about it. Writing assessments which are not interpreted for their curricular implications and which are not followed by inservice educational programs to improve instruction waste teachers' and students' time, and taxpayers' money.

A Critique of Current Assessments

Teachers might better tolerate the dearth of inservice education as a major component of assessment programs if they were persuaded that current programs truly assessed students' ability to write and truly yielded accurate information. But the truth of the matter is that all assessment programs are highly inadequate at present for reasons previously cited: they rely too heavily on editing skills, and when they do include students' writing, they sample under highly artificial conditions too few pieces in too few modes in far too short a period of time. The consequence is that inferences and conclusions based upon assessment findings, though they might provide sensational information for the media, make for shoddy research.

To be adequate, assessments conducted in the secondary schools would need to draw samples of writing from across the spectrum of rhetorical modes or function categories, from narration, description, classification,

and evaluation, for example, or from expressive, transactional, and poetic discourse. Each mode or category requires the use of discrete skills, and the student capable of writing a poignant account of a personal experience may not command the rhetorical skills to write a closely reasoned argument on a contemporary issue. Since the quality of writing of even the brightest student, regardless of grade in school, may vary considerably from day to day, the modes or categories of writing to be assessed would need to be sampled across a span of time, perhaps weeks. Moreover, since the quality of a student's response to similar compositional tasks may likewise vary considerably from one assignment to the next, a full assessment would require at least two samples to be evaluated for each category or mode. Finally, the assignments would need to be so constructed as to minimize their artificiality; students would need to be given adequate time to engage themselves fully in the processes of prewriting, writing, and revising; and they would need to write in their own classrooms, with full opportunity to ask questions of both their teacher and their classmates and to consult dictionaries if they wished. In short, they should be given the chance, as much as conditions permit, to behave the way that those adults who write regularly behave.

I fully realize the fiscal implications of what I am proposing, but anything less deludes the public into believing that we know how well, or how poorly, students in this nation write. We do not, for we have not as yet created the stimuli, the processes, and the conditions that would permit us to know.

Rather than evaluating a range of papers written at different times in different modes or categories, most large-scale assessment programs attempt to economize by requiring students of a given grade to respond to a single writing task. When that writing task differs in kind from one

grade level to another, resultant data are quickly misinterpreted by the public. I know of one state in which fifth graders were asked to pretend they were newscasters describing an event depicted in an illustration accompanying the assignment; whereas, ninth graders were asked to write a letter to a school administrator arguing for retaining in the curriculum a program threatened with budgetary elimination. In short, the first assignment called for descriptive/narrative reportage, while the second required a logically organized persuasive argument. When the papers for both grades were read holistically by trained readers employing a scale of 1 to 4 (4 being highest) the fifth-graders received far many more grades of 3 and 4 than did the ninth graders. The inference immediately drawn by those unfamiliar with the disparate rhetorical demands of the two assignments was that between grades five and nine, the writing abilities of students precipitously decline, the consequence, no doubt, of faulty instruction. Such inferences are almost impossible to head off at the press, and once printed and broadcast, can do irreparable harm to teachers' morale.

Although those conducting current assessments of writing are cognizant of the severe limitations of their programs and although they try to warn the public--often it seems in a whisper--of those limitations, the caveats get buried in the glare of publicity that follows the release of data. One finds, for example, on page 4 of NAEP's Writing Mechanics, 1969-1974 the following statement under the heading "The Limitations of the Study":

> This report describes only one facet of the national assessment of writing. It deals with one particular writing task performed under one particular set of circumstances and in no way constitutes a definitive study of writing in the largest sense of that word. The students who participated in this exercise were told to use their best writing but were not told that their papers would be examined for mechanical correctness.

Those familiar with NAEP processes were aware that, aside from Puritan compunctions, students had no real reason to do their best writing, nor did they always have adequate time to proofread and carefully revise their papers. Although the brief statement of "the limitations of the study" is easily forgotten in a report of 59 pages, it may have salved some consciences at NAEP headquarters. However, when NAEP two years later published What Students Know and Can Do (1977), it failed to indicate that the studies that gave rise to its sweeping generalizations about students' writing abilities were in the least limited.

Recommendations for Change

To summarize, I find large-scale assessments of writing to be inadequate, costly in terms of both money and time, ineffectual in improving instruction, misleading to the public, and detrimental to teachers. Consequently, I believe they should be abolished. In their stead, I would propose local assessments of writing, assessments tailored to the particular instructional goals of local communities. Inservice-education for teachers of English language arts would be an integral component of the assessment programs I envision: education in how to design sound composition assignments, in how to evaluate papers judiciously, in how to diagnose writing problems, in how best to assist those with severe difficulties, in how to assess locally students' longitudinal progress toward achieving writing proficiency. In Measuring Growth in English (NCTE, 1974), Paul Diederich has described one method by which such local assessments of writing might be conducted. There are undoubedtly others, including sampling from portfolios of students' writing and conducting "blind" evaluations of papers written on comparable topics early and late in the school year. In fashioning local assessments

and the inservice educational programs that would accompany them, those who have been engaged in state and national assessments and who have learned from their limitations might serve as invaluable advisors.

Carefully conceived, local assessments should be far less bureaucratic in nature than either state or national assessments, far more attentive to the educational problems of the handicapped and of minorities, including bilingual students, far more attuned to local values and local concerns. Carefully conceived, they should negate the criticisms that Merle McClung of the Center for Law and Education at Cambridge has advanced against competency tests, namely, that they often lack instructional and curricular validity (see, for example, McClung's "Competency Testing: Potential for Discrimination," Clearinghouse Review, September 1977). Carefully conceived, they should obviate the criticism that Paul Olson directs at NAEP in A View of Power..., namely that it is "an assessment that, at best, asks powerless communities to assess themselves in terms provided by the powerful." Olson defines the powerless as "those persons who would, under law, be clients of the ESEA Title I program, the Bilingual Education Act, the Indian Education Act, and other similar programs," as well as "those persons who would not be comfortable at T.S. Eliot's Cocktail Party." (p. 1) Carefully conceived, local assessments should help stop the homogenization of education; should help return control of the curriculum to parents, teachers, and students; should lead to better instruction from teachers and better writing from those in their charge.

And with the latter goal, at least, we should have no disagreement.

BIBLIOGRAPHY

A Citizen's Introduction to Minimum Competency Programs for Students. Columbia, South Carolina: Southeastern Public Education Program, American Friends Service Committee, February 1978.

Basic Skills Assessment Package. Princeton, NJ: Educational Testing Service, 1977.

Comptroller General of the United States. The National Assessment of Educational Progress: Its Results Need to Be Made More Useful. Washington, D.C.: National Center for Education Statistics, Department of Health, Education and Welfare, July 20, 1976.

Council-Grams, Urbana, Illinois: National Council of Teachers of English, March, 1980.

Frahm, Robert and Jimmie Covington. What's Happening in Minimum Competency Testing. Bloomington, Indiana: Phi Delta Kappa, 1979.

Huntley, Renee M. What's Really Basic in the Language Arts? Iowa City, Iowa: American Council on Testing Publications, 1977.

Irmscher, William F. Teaching Expository Writing. New York: Holt, Rinehart and Winston, 1979.

McClung, Merle Steven. "Competency Testing: Potential for Discrimination," Clearinghouse Review, August/September 1977.

McClung, Merle Steven and Diana Pullin. "Competency Testing and Handicapped Students," Clearinghouse Review, March 1978.

Mellon, John C. National Assessment and the Teaching of English. Urbana, Illinois: National Council of Teachers of English, 1975.

Odell, Lee. "Defining and Assessing Competency in Writing." Unpublished.

Olson, Paul. A View of Power: Four Essays on the National Assessment of Educational Progress. Grand Forks, North Dakota: University of North Dakota, December 1976.

What Students Know and Can Do. Denver: National Assessment of Educational Progress, March 1977.

Williams, John D. Testing and the Testing Industry: A Third View. Grand Forks, North Dakota: University of North Dakota, December 1975.

Writing Mechanics 1969-1974. Denver: National Assessment of Education Progress, October 1975.

ALTERNATIVE METHODS OF WRITING ASSESSMENT

Richard J. Stiggins
Clearinghouse for Applied Performance Testing

There are two viable approaches to the assessment of writing proficiency. One is termed the direct method. It relies on actual samples of student writing to judge writing proficiency. The second method is the indirect method. It relies on objective tests. The following presentation provides a detailed analysis of the similarities and dissimilarities between these two approaches of assessing writing proficiency.

Although the recent development of various new direct measurement procedures was sorely needed and has added significantly to the array of measurement approaches available to the writing instructor, one effect of this developmental emphasis has been to inappropriately reduce the use of indirect writing assessment methods in instances where they might profitably be used. It is my opinion that both direct and indirect assessment methods should be given consideration. The analysis presented herein will describe the basic characteristics of each approach and will lead to the conclusion that each can be used to advantage in certain educational assessment contexts.

Direct and indirect assessment methods will be compared on the basis of seven specific criteria--educational context, assessment focus, practical consideration, test item characteristics, response factors, test scoring methods and procedures for judging test quality. The first factor comparison considers how each assessment method relates to the educational context, and what kinds of educational decisions each method can be expected to facilitate.

The second criterion of comparison is the focus of the assessment. This deals with the component of writing tested, the conditions under which that assessment takes place and the kinds of traits that can be measured by direct and indirect writing assessment instruments.

The third criterion focuses on practical considerations, including the kinds of attitudes that test users need to have for successful assessment, options for test acquisition, the kind of expertise required in conducting the assessment and an analysis of the respective costs associated with conducting direct and indirect assessments.

The fourth criterion considers test exercises or test items and compares the type of exercises administered, the flexibility of those exercises, the fidelity (or relationship of the exercises to real life circumstances), and the manner in which the test user exercises control over the kinds of skills tested.

The fifth factor is the kind of response that is elicited by the respective test exercises. This covers the examinee's task in the assessment, the examinee's relationship to the text of the assessment and the fidelity of the response elicited by the assessment.

The sixth factor deals with test scoring procedures and covers alternative methods of scoring the assessment, methods of scaling

performance, objectivity of the scoring process, the degree of precision that can be obtained in scoring and score reporting alternatives.

The seventh criterion of comparison is methods used to judge test quality, including methods of estimating score reliability and score validity.

And, finally, the comparison of direct and indirect methods concludes with a summary of the major advantages and disadvantages of each, and a review of the specific roles that direct and indirect assessment methods can play in eight specific educational decisions.

A series of charts appended to the paper also concisely reviews the seven point analysis of similarities and differences between direct and indirect methods.

Criterion 1. Relationship to the Assessment Context

Although fundamentally different in form, both direct and indirect writing assessment methods can be equally useful in a variety of educational assessment contexts: Each provides an effective way of measuring a student's ability to use standard written English.

However, despite the similarity in test purpose, the tests and the nature of the information that can be gained from each differs substantially. In the case of direct assessment, the examinee is given prompts which elicit writing samples that are evaluated according to prespecified criteria. In indirect methods, the examinee responds to a series of objective test items that call for the examinee to evaluate and judge the appropriate use of language in a series of test items. Each of these testing methods requires that the examinee demonstrate the ability to apply previously acquired knowledge about language usage.

Each measurement approach provides information that is useful in helping the educator make a variety of specific educational decisions. For example, results can be used to identify the best instructional alternatives (instructional management) for a student, in screening students into special programs, or for evaluating programs. Instructional management decisions, which include diagnosing student strengths and weaknesses, placing students in proper writing instruction programs, and helping students to make vocational and educational guidance decisions, can be based on either measure. Direct measures of writing proficiency, using writing samples, can be a valuable tool in making these decisions, so long as resources are available to conduct relatively detailed analyses of the proficiencies demonstrated by students in the writing samples. Indirect writing assessment methods also serve in instructional management contexts if criterion referenced tests are used. Such tests break students' overall writing performance down into component parts, allowing for a detailed analysis of developed and undeveloped skills.

Both indirect and direct methods can also serve in screening examinees with regard to minimum essential skills. If the scoring criteria used to rate writing samples are designed to assess predefined essential skills, then writing samples can be used to select students and certify minimal competencies. The same is true of indirect measures, if the items included on the test are items that test the skills essential for certification of competence.

Program planning decisions (or program evaluation) can also be based on direct or indirect methods. In each case, if the scoring criteria (direct measures) or the test items included on the test (indirect) are directly related to (or are representative of) the intended outcomes of a particular educational program, then those tests can serve as a basis for

evaluating how well the program under investigation is, in fact, achieving its intended outcomes. In short, both assessment methods can serve program evaluation purposes if they have curricular validity.

To summarize, both direct and indirect methods have the same underlying purpose--that of measuring the ability to use standard written English. Though obviously each method uses different measurement strategies, both are useful in a variety of important educational assessment contexts. This point will be amplified later.

Criterion 2. Focus of the Assessment

Direct and indirect writing assessment methods focus on different components of the writing process. In the case of direct assessment, the component of writing tested is actual composition skill. In indirect assessment, the component tested is proficiency with some of the prerequisites of effective writing--grammar, punctuation, sentence construction, organization, etc. In direct assessment, the information gathered provides necessary and sufficient information for drawing conclusions regarding the student's actual writing proficiency. Indirect assessment on the other hand, covers some of the necessary but not sufficient conditions for evaluating the student's writing proficiency.

Reviewing the kinds of traits that are measured in the two approaches reveal some of the differences in assessment focus. Here are some examples distilled from the 18 statewide writing assessments currently underway in the United States:

DIRECT ASSESSMENT	INDIRECT ASSESSMENT
Usage	Punctuation
Sentence sense	Grammar
Expression of feeling	Placing modifiers
Persuasiveness	Determining idea relevance
Organization	Diction
Format	Style
Cohesiveness	Transition

Revision skills	Logic
Transition	Organization
Overall writing proficiency	Overall usage proficiency
	Sentence structure

Criterion 3. Practical Considerations of Testing

In the development and implementation of a writing assessment, there are several very important practical factors that have to be taken into account to ensure quality assessment. Considerations of these factors reveal additional differences between direct and indirect assessment methods.

Key Attitudes. The first practical considerations essential for the success of the assessment are the attitudes of users of the writing test. In the case of direct assessment procedures, assessors and users of the data must be willing to invest time, money and effort to conduct a writing assessment that calls for using complex testing procedures (outlined below). However, in the case of indirect assessment, the key beginning attitude is a willingness to accept a proxy measure, or a measure that is only an approximation of the actual trait which is the focus of the assessment. If, in either case, the appropriate attitudes are present, the assessment will most probably have its desired impact. If either is missing, problems can be anticipated.

Test Acquisition and Development. Another practical consideration in testing is the manner of acquiring the test. In both direct and indirect assessment, the examiner has two choices, (a) select an already existing test, or (b) construct a new test.

If one is considering the use of a direct assessment procedure and anticipating selecting a previously developed set of exercises and scoring criteria, then the kind of expertise required to plan the assessment will include (1) technical expertise in the area of writing to specify the

writing skills to be assessed; (2) test evaluation skills, to investigate available writing sample options and select a test that matches the identified component skills; (3) organizational skills to set up, administer, score and report the results of the assessment.

In developing a new direct assessment instrument, which involves creating a new set of exercises and criteria for scoring, organization and technical writing expertise will again be necessary. However, in this case, psychometric expertise will also be required to evaluate and refine the validity and reliability of the assessment procedures.

Selecting a previously developed indirect assessment instrument calls for expertise very much like that required to select a direct measure. This includes technical writing expertise to identify the skills to be assessed and test evaluation expertise to review, evaluate and select a test that matches the desired writing skill criteria.

Finally, in developing a new indirect assessment instrument, it is necessary to have (1) technical expertise in writing to plan the assessment; (2) item writing or item selection capability to construct the component parts of the new test; (3) organizational skills to pilot test, analyze and select the new items to appear on the test; and (4) psychometric expertise to evaluate the reliability and the validity of the assessment instrument.

In general, then, the development of new instruments for either testing approach requires substantially more expertise and staff time than does the use of already existing assessment instruments.

Testing Costs. The cost of assessment is perhaps the single most important practical consideration in deciding what assessment approach should be used in a particular setting. There are three kinds of costs to be considered: developmental costs, test administration costs, and test

scoring costs. The cost factors for direct or indirect assessment vary considerably and depend, to a great extent, on whether new instrumentation is to be developed or already existing instrumentation is to be used.

To illustrate, the developmental costs that can be anticipated in the selection of a previously developed direct assessment method include:

1. Time required to plan the assessment
2. Time required to secure, review, evaluate and select existing exercises and a scoring guide
3. Cost of producing all necessary test materials

Costs that can be anticipated in developing a new direct assessment instrument are somewhat more extensive and will include:

1. Cost of time associated with planning the assessment
2. Time associated with the development of exercises
3. Time associated with development of a scoring guide
4. Time and administrative costs associated with field testing the exercises and scoring procedures
5. Costs of producing all necessary test materials

Developmental costs that will be incurred selecting an already existing indirect assessment instrument are similar to those in the direct assessment and include:

1. Time associated with planning the assessment
2. Time necessary to review, evaluate and select the test
3. Cost of purchasing all necessary test materials

And, finally, costs associated with the development of a new indirect assessment instrument--potentially the most costly option available--include:

1. Time required to plan the assessment
2. Costs associated with item writing (or selection of an item pool, if available)
3. Cost of pre-testing the items, analyzing the data and assembling the test
4. Cost associated with norming the test (if necessary)
5. Cost associated with the production of all necessary test materials

The test administration costs will generally remain constant regardless of the assessment approach used or whether or not new instruments are used. Those costs include budgeting adequate time to plan

the test administration, select and train test administrators, coordinate the materials distribution, administer the test and coordinate the collection of test materials.

Test scoring costs differ substantially with the two methods. For example, scoring costs of direct assessment--whether with new or existing instruments--includes this rather lengthy and costly list of items:

1. Cost of time associated with planning relatively complex scoring procedures
2. Costs of facilities use
3. Costs associated with selecting, recruiting and training scorers
4. Cost of staff time required to read and rate the writing samples
5. Costs associated with the evaluation of the reliability of the scores
6. Costs associated with processing the scores for reporting

On the other hand, the costs associated with scoring an indirect assessment (objective test) are less extensive and include costs for computer time to machine score the tests or personnel time to hand score. In developing a new indirect instrument, an added cost will be incurred in evaluating the reliability of the newly developed test. With both direct and indirect assessment, however, there will be costs associated with processing the scores for reporting.

A final testing cost factor which clearly differentiates direct and indirect assessment methods is that of test reuse. In either case, if security is not breached or if test security is not an issue, previously developed test exercises can be reused. However, this reuse of direct and indirect tests has a differential impact on overall testing costs. With direct assessment, test development costs are low, while scoring costs are high. The opposite is true of indirect tests; test development costs are high, while scoring costs are low. With the reuse of either, the test development costs can be amortized with multiple use of the test; scoring costs, however, recur with every use. With indirect assessment, the high

cost of test construction can be amortized over repeated administrations and the recurring scoring costs are very low. However, with direct assessment, although the initial developmental costs are low, the scoring costs remain high with each use. Therefore, depending on the initial test development costs, direct assessment could result in significantly higher testing costs than indirect, especially if the test is used numerous times. An informed decision regarding which approach to use will include a cost/benefit analysis that takes these factors into account.

Summary. To summarize practical considerations in the direct/indirect comparison, the key beginning attitude in the case of direct assessment is the willingness to invest in what will obviously be a complex assessment process, while the key beginning attitude for the indirect assessment alternative is willingness to accept a proxy measure. In both cases, it is possible to either select or construct an assessment instrument, this carries with it implications for the kind of expertise required to conduct the assessment and cost incurred. The principal costs associated with direct assessment are likely to be scoring costs, while the principal costs associated with indirect assessment are likely to be test development costs.

Criterion 4. Factors Related to Test Exercises

There are some fundamental differences in the kinds of test exercises used in direct and indirect writing assessment instruments. First of all, the form of the basic exercise differs. In the direct method, prompts are used to identify for the examinee (a) the form of writing to be produced, (b) the audience to be addressed, and (c) the purpose for the writing. In the case of indirect assessment, various written items using a multiple

response format are used. Examples include various interlinear forms, as well as sentence combining items and other kinds of formats.

As a result of the format and exercise differences, direct assessment exercises are considerably more flexible than indirect. With direct assessment stimulus materials can be auditory or visual and can focus on an infinite array of real life circumstances. Indirect test items, on the other hand, are generally constrained to the particular written item format required by the form of test in use. As a result, the tests differ in fidelity, or the extent to which they approximate "real world" writing. Direct assessment exercises can be made to closely approximate "real world" writing requirements, while objective test items are somewhat more artificial.

The amount of control that can be exercised over the skills tested is another difference between direct and indirect exercises. In direct assessment, the examiner controls the kinds of skills tested through selecting or developing prompts that identify the form of writing to be demonstrated, the audience to be addressed and the purpose for the writing. It is the examinee's responsibility to demonstrate the required proficiency within the framework provided by the prompt.

However, in the case of the objective test items, control over the kinds of skills tested is exercised through the selection of items included in the test. Careful selection of sampled items can give the examiner a very high degree of control over the specific skills tested.

Consequently, in the direct case, there is some probability of sampling error. For example, in the area of mechanics of language usage, it is possible for the less-than-proficient examinees who are composing a writing sample to avoid unfamiliar or difficult sentence constructions. Thus, the examiner might erroneously conclude that examinees are competent

in certain usage areas, when, in fact, they are not. The same is not true in the case of indirect assessment where sampling error is controlled by forcing the examinee to deal with all relevant instances of usage and demonstrate mastery or non-mastery of specific elements of English mechanics. In effect, they have no choice in responding to items and, therefore, cannot escape from demonstrating their lack of proficiency.

Consequently, with regard to test exercises, one maximizes the authenticity and flexibility of the assessment by using direct assessment exercises, but, in doing so, sacrifices control over the kinds of skills tested. With indirect assessment, one sacrifices fidelity to the real world in exchange for control over the specific skills tested.

Criterion 5. Response Factors

As we examine the responses of examinees in direct and indirect assessment, we again see fundamental differences between the two. The examinee's task in the indirect case is to read, evaluate and select the appropriate usage. With direct assessment, on the other hand, the examinee's task is to plan, read, and revise the text. In the direct assessment, the examinee constructs the text; in the indirect case, the examinee evaluates the text.

By planning, writing and revising the text in the direct method, the examinee is replicating the kind of writing behavior that exists in the non-school world. The fidelity of the response is high. However, in indirect assessment, the process of reading, evaluating and selecting does not closely approximate real world writing circumstances; consequently, the fidelity of the response is low.

Criterion 6. Factors Related to Test Scoring

As was suggested in considering testing costs, there are substantial scoring differences between direct and indirect assessment methods. First of all, with direct assessment there are various methods of scoring writing samples. These include holistic, analytical, primary trait, cohesiveness, syntactic sophistication and others and scoring methods can be adapted to any particular local educational assessment context in which writing skill is a key consideration. In the indirect case, on the other hand, there are just two options. One can score the entire test or break the test items down into subscores.

The method of scaling or mapping performance on a numerical scale, in direct assessment, calls for developing performance rating scales or checklists of demonstrated proficiency. The scaling method in indirect assessment involves counting items answered correctly.

The objectivity of the direct assessment approach relates directly to the specificity of the scoring criteria and the quality of the rater training procedures. In the indirect case, on the other hand, the objectivity of the test is a function of item writer capability (the quality of the items) and item selection procedures.

However, though differences exist in scoring methods, scaling, and determination of objectivity, there are score reporting similarities between the two. For example, each can be used to generate either holistic or analytical scores. That is, each can provide an overall assessment or an analysis or diagnosis of the component parts of demonstrated writing proficiency. Also, in either case, norm referenced or criterion referenced score interpretation is possible. That is, examinees can be compared to each other or to a common standard of acceptable performance.

Criterion 7. Judging Test Quality

The factors commonly considered in judging the psychometric adequacy of a test are reliability and validity. The relationship between direct and indirect writing measures is discussed in this section in terms of these two factors.

Reliability. Like any test, a writing test must yield dependable or reliable scores to be useful; in both methods of writing assessment, reliability takes four forms. Test scores must be stable over time, across parallel test forms, across different parts of the same test, and across raters. Otherwise, the examiner will not know if the score can be relied on as an accurate estimate of the examinee's real proficiency.

The following brief examples illustrate the four aspects of reliability. Suppose a direct writing assessment were administered to the same students twice, the second administration following a two- to three-week interval. And suppose that even though no writing instruction had taken place, the scores obtained the second time were totally different from those achieved the first time--for nearly every examinee. The examiner would not know which score (if either) to depend on as the true reflection of the student's proficiency. Or, suppose two writing exercises were developed to measure exactly the same skills and yet when both were administered to a student, the exercises resulted in totally different estimates of proficiency. Again, the examiner would not know which score was the better indicator of proficiency. Or what if ostensibly equivalent forms of a performance test resulted in totally different estimates of proficiency. The examiner would not know which form to rely on. Or from a fourth perspective, suppose two judges evaluated the same performance and drew totally different conclusions

regarding proficiency. In this case, the examiner would not know which judge to rely on.

When scores on tests are unstable over time, differ considerably across ostensibly equivalent exercises or test forms and/or differ substantially across evaluations of proficiency by independent judges, there is reason to question the usefulness of the assessment procedures. When these differences occur, the examinee's score might indicate being influenced by administration time, the particular exercise or test that happened to be used, or the rater that happened to evaluate the examinee's response--all factors independent of the examinee's real proficiency. These are undesirable determinants of test scores. However, when the writing assessment procedures employed yield scores that are stable over time, across exercises, test forms and independent evaluators, these scores can be confidently used for educational decision making.

The fundamental difference between direct and indirect writing assessment procedures is that consistency across raters is not an issue in the case of objective instruments. However, consistency over time, across items, and across test forms remains relevant for both assessment approaches.

Another way of viewing the issue of test score dependability is from the perspective of potential sources of unreliability or potential reasons for inaccurately measuring a student's writing skill. These reasons differ between direct and indirect approaches. In direct assessment, inaccurate scores can arise from (1) poor exercises (e.g., insufficient focus), (2) poor test administration procedures or environment, or (3) poor scoring procedures (rater leniency, halo effects, tendency toward middle rating, etc.). In the case of indirect assessment, poor test items and test administration procedures can also give rise to inaccurate

scores. However, in place of poor scoring procedures, which are not a factor in the indirect case, guessing also becomes a possible source of error.

Another dimension of score reliability that often becomes particularly important in large scale writing assessments is score scale consistency across a span of time (say, from one year to the next). Procedures are often needed to ensure that a score given one year is directly comparable in meaning to the same score the next year, particularly in instances where, say, a high school diploma rests in the balance. In the case of indirect assessment, this equating of scores is accomplished via complex statistical procedures. In direct assessment, equivalence is attained through careful rater training to be sure that they understand and apply the same rating criteria and standards each year.

Validity. There are two important considerations related to the validity of writing tests. The first is content validity, or the extent to which the objective test items or writing exercises and scoring procedures do, in fact, represent the kinds of skills that are the intended focus of the test. In both direct and indirect measures, this kind of validity is a matter of expert judgment: Test items or exercises and scoring guides (whichever is used) must be reviewed by independent writing specialists to ensure their appropriateness.

The second type of validity--criterion-related validity--must be considered from two perspectives: internal validity and external validity. Internal validity is verified by correlating writing test scores with other simultaneously administered measures of writing proficiency. In the case of indirect measures, internal validity must always be verified by showing a high correlation between the objective test score and score attained through the use of actual writing samples

(direct measures). Internal validity of direct measures can be verified by correlating scores with other writing performance indicators such as courses taken, grades received or other direct measures of writing skill.

The external validity of a writing test reflects the extent to which the scores are correlated with indicators of writing proficiency at other times under other circumstances. For example, the scores of either direct or indirect measures can be shown to be valid from an external perspective if they predict subsequent English course grades.

Summary. In judging the psychometric adequacy of a writing test, the reliability and validity considerations for direct and indirect measures are quite similar. In the case of direct measures, score stability over time, across exercises, across test forms, and across raters are important. All of these types of score consistency except for stability across raters, are also relevant for indirect measures. In both cases, sources of inaccurate scores include poor test items and poor administration. Unique sources of score inaccuracy include guessing, in indirect measures, and poor scoring in direct measures. Score scale equivalance from one administration to the next is maintained via statistical methods for indirect measures and rater training for direct. Content validity is relevant to both types of writing tests and is verified in both via expert judgment. And criterion-related validity is also important in both cases and is verified in terms of correlations with other indicators of writing proficiency.

Summary of the Comparison

The differences and similarities between direct and indirect methods of writing proficiency can be summarized by reviewing the major points of comparison.

Advantages and Disadvantages. First, in direct assessment, carefully trained raters are the key to high test quality. Raters must internalize the criteria and standards of acceptable performance in order to maintain test reliability and validity. In indirect assessment, the key determiner of test quality is the test item writer and the test developer. A reliable and valid writing test will result if (1) the item writer internalizes the skills to be tested and reflects those skills in the objective test items, and (2) the test developer appropriately selects samples from among those items to assess the desired domain of writing skills.

Second, a detailed breakdown of assessment costs indicated that, in direct assessment, scoring is the major cost factor. With indirect assessment the major cost is associated with the cost of test development.

The major advantages of the direct assessment option are (1) the potential richness of the result of the assessment--a sample of writing contains a great deal of information about the examinee's real writing proficiency; (2) high fidelity of the stimulus and response; (3) the flexibility or adaptability of the exercise to a variety of relevant real world writing circumstances; (4) the positive test user's attitude that result from the high face validity of writing samples; and (5) the relatively low test development costs associated with the use of writing samples.

The major advantages associated with the indirect assessment option, on the other hand, are: (1) high score reliability; (2) relatively low test scoring costs; and, (3) high degree of control over the nature of the skills tested.

The potential disadvantages of the direct method include: (1) the high cost of scoring; and, (2) the potential lack of standardization in the nature of proficiencies assessed across examinees.

The potential limitations of the indirect method are: (1) the lack of fidelity or correspondence to real world writing tasks; (2) heavy reliance of the assessment method on reading skills rather than writing proficiency; and in many cases (3) lack of test face validity in the objective measure.

Uses of Writing Tests. Given these advantages and disadvantages, it becomes possible to identify the conditions under which direct and indirect assessments can be used in the various educational decision making contexts outlined in the introduction. For example, in the instructional management decision domain, there are three particular kinds of decisions in which both direct and indirect writing assessments can play a significant role. These three are the diagnosis of strengths and weaknesses, course placement, and guidance decisions.

In diagnosing a student's strengths and weaknesses, direct assessments are best used with relatively small groups in which a teacher can conduct a detailed analysis of writing samples to determine the particular strengths and weaknesses of students. In larger groups of students, direct assessment may be less useful for diagnosis of strengths and weaknesses because of the high cost of analytical scoring incurred in this setting.

If indirect writing tests are criterion referenced, they can also be used for the diagnosis of strengths and weaknesses. That is, the tests must include items that are specifically referenced to important elements of the writing process, and yield scores that reflect the student's mastery of this particular writing component. If these conditions are satisfied, an objective test can be quite diagnostic.

In both course placement and guidance decisions, direct assessments can play a role if relatively low-cost scoring procedures are selected. In most cases, that means using the holistic scoring process and arriving at an overall score regarding the level of the student's development as it relates to the performance of other students in the group. This kind of comparative test score can help students and teachers make decisions about the next most appropriate course of instruction for the student and aid in educational and vocational planning.

Indirect assessment methods can also be used in placement and guidance decisions and, in fact, are more commonly used in these contexts than direct assessment methods. It is not uncommon to see a published, norm referenced test of language usage proficiency being used to rank students' usage skills for placement into various developmental writing courses. It is also common to use these tests to help students make educational and vocational decisions for entering college programs in which writing proficiency is a significant factor in success.

Writing tests can also play a role in the screening of examinees. Two examples of screening decisions are (1) selecting students for a particular program (remedial or advanced) and, (2) certifying minimum competencies. Direct writing assessment instruments can be used for the selection process and are, in fact, the best choice when writing proficiency is either the only or the key selection criterion. For example, direct measures of writing proficiency are most appropriate when a test is being used to select students to receive four-year scholarships for a college writing program. Under these circumstances, the test with maximum fidelity--a direct assessment measure--would be the test of choice. However, indirect measures of writing proficiency also can be used in selection decisions and are acceptable whenever writing

proficiency is one of many (rather than the only) selection criterion. With respect to certification decisions, direct measures of writing proficiency are useful whenever the scoring scheme has been carefully developed and the students' performance is rated on the competencies that are considered the minimum acceptable competencies. Indirect measures of writing proficiency can also be useful in making minimum competency decisions if the test items have been selected carefully to represent minimum acceptable skills.

The third set of educational assessment contexts in which writing tests can play a role are decisions having to do with program planning or program evaluation. There are three examples of these kinds of decisions: (1) survey assessment, (2) formative program evaluation and, (3) summative program evaluation. In survey assessments such as statewide assessment programs, tests are administered to a representative group of students for drawing general conclusions about the level of proficiency attained by that student group. Since these surveys typically include such large samples of students, the most efficient direct assessment method should carry with it low-cost scoring procedures. Holistic scoring, for example, may play a role here. However, in the instance of large-scale assessments, it might be useful to take advantage of the high degree of scoring efficiency associated with the indirect option. To appropriately use either method in formative or summative program evaluation requires that the intended outcomes of the course under investigation be reflected in the direct assessment scoring criteria and in the items included in the indirect assessment. That is, either direct or indirect assessment methods may be useful if they really test what is being taught.

On the basis of this analysis of the similarities and differences between direct and indirect assessments of writing proficiency, it is apparent that each has a valuable role to play in the assessment of student writing proficiency. The preceding presentation has been an attempt to outline the various conditions under which each method can and should be used. It is inappropriate to conclude that one of the two options is acceptable and the other unacceptable; they're both acceptable, but their usefulness varies according to the educational assessment context and the decisions to be made. Those planning to assess writing proficiency might profitably conduct a thorough analysis of their particular educational and assessment needs and give careful consideration to both of the viable assessment options available to them--direct and indirect writing assessment--before making a final selection.

COMPARISON OF

DIRECT AND INDIRECT WRITING ASSESSMENT METHODS

BASIS OF COMPARISON	DIRECT ASSESSMENT	INDIRECT ASSESSMENT
1. Relationship to assessment context		
A. Reason for testing	Measure ability to use standard written English	Measure ability to use standard written English
B. Kind of test administered	Prompts eliciting writing samples	Objective test
C. Decisions tests can help with		
(1) instructional management	Requires detailed analysis of writing samples	Requires criterion-referenced tests
(2) student screening	Scoring criteria must cover essential skills	Items must test essential skills
(3) program planning	Scoring criteria must relate to program outcomes	Items must test intended outcomes
2. Focus of the assessment		
A. Component of writing tested	Composition skill	Proficiency with some prerequisites of effective writing
B. Conditions of assessment	Covers the necessary and sufficient conditions for evaluating writing skill	Covers the necessary but not sufficient conditions for evaluating writing skill
C. Examples of traits measured	Usage, sentence sense, expression of feeling, persuasiveness, organization, format, cohesiveness, revision skills, transition, overall writing proficiency	Punctuation, grammar, sentence structure, placing modifiers, determining idea relevance, diction, style, transition, logic, organization
3. Practical considerations in testing		
A. Crucial beginning attitude	Willingness to invest time, money and effort in conducting sound assessment	Willingness to accept results on a proxy measure
B. Options for test acquisition	(1) select previously developed exercises & scoring criteria (2) construct new test	(1) select published test (2) construct objective test
C. Expertise required	1. organizational 2. technical writing (1) 1. organizational 2. technical writing 3. psychometric	1. technical writing 2. test evaluation (2) 1. technical writing 2. item writing (or selecting) skills 3. organizational (test dev.) 4. psychometric

COMPARISON OF ASSESSMENT METHODS

BASIS OF COMPARISON	DIRECT ASSESSMENT		INDIRECT ASSESSMENT	
	(select)	(construct)	(select)	(construct)
D. Testing costs				
(1) Developmental costs	plan assessment secure, review, evaluate & select exercises & scoring guide prepare all test materials	plan assessment develop exercises develop scoring guide field test exercises & scoring prepare all test materials	plan assessment review, evaluate & select test purchase test materials	plan assessment write items (or select from pool) pretest items assemble test norm (if necessary) prepare test materials
(2) Test administration costs	plan administration prepare instructions & train administrators coordinate materials distribution administer (can include revision) coordinate materials collection	(same)	(same)	(same)
(3) Test scoring costs	plan scoring procedures arrange for facilities select scorers train scorers read papers evaluate reliability process scores for reporting	(same)	score test by hand or machine	score test evaluate test (items & reliability) process scores for reporting
(4) Implication of test reuse	Exercise and scoring guide can be reused if security is maintained or is not an issue Relatively high scoring costs constant with each use		Test can be reused depending on security Significantly lower scoring cost constant with each use	
4. Factors related to test exercises				
A. Exercise form	Prompts that identify form of writing, audience and purpose for writing		Various written item forms with multiple response options (e.g., interlinear, multiple-choice)	
B. Flexibility	Infinitely flexible: stimulus materials can be auditory or visual; prompt can focus on wide variety of writing circumstances		Less flexible: items generally presented in written form	

COMPARISON OF ASSESSMENT METHODS

BASIS OF COMPARISON	DIRECT ASSESSMENT	INDIRECT ASSESSMENT
C. Fidelity	Can very closely approximate real world tasks	Under any circumstances, somewhat artificial
D. Source of control over skills tested	Controlled through exercise specificity: form, audience, purpose	Controlled through selection of test items
	Some sampling error possible (e.g., usage)	Sampling error controlled via item selection
5. Factors related to response		
A. Examinee's task	Plan, write, revise	Read, evaluate, select
B. Examinee's relationship to text	Examinee constructs text	Examinee evaluates text
C. Fidelity	Very close approximation to real world behavior	Generally artificial
6. Factors related to scoring		
A. Alternative methods	Holistic, analytical, primary trait, cohesiveness, syntactic sophistication and others	Score, subscores
B. Method of scaling	Rating scale(s) or checklist	Item count
C. Level of objectivity	Function of trait specificity and rater training	Function of item writer capability and item selection
D. Score precision	Holistic or analytical	Holistic or analytical
E. Score reference	Criterion or norm reference	Criterion or norm reference
7. Judging test quality		
A. Score reliability factors		
(1) consistence required across	Exercises, test forms, time and raters	Items, test forms, time
(2) possible reasons for inaccurate measurement	1. Poor exercises: insufficient focus exercises don't measure skills 2. Poor test administration 3. Scoring procedures: leniency mid-scale tendency halo effect logical or proximity effects rater self-perception effects	1. Poor items: unfocused items items that fail to match skill biased items 2. Poor test administration 3. Guessing

COMPARISON OF ASSESSMENT METHODS

BASIS OF COMPARISON	DIRECT ASSESSMENT	INDIRECT ASSESSMENT
(3) Score scale consistency (equating)	Managed through rater training	Managed through statistical methods
B. Validity factors		
(1) content validity	Exercise fidelity evaluated via user judgment	Item representativeness evaluated via user judgment
(2) criterion-related validity	Correlation with other performance indicators	Correlation with direct measures
8. Summary of methods		
A. Key to high test quality	Test scorer	Test item developer
B. Major cost factor	Scoring	Test development
C. Major advantages	Potential richness of result High fidelity of stimulus and response Exercise flexibility Potential for positive user attitudes Relatively low test development costs	High score reliability Relative low scoring cost High degree of control over skills tested
D. Potential disadvantages	High cost of scoring Potential lack of standardization	Lack of fidelity Reliance on reading Lack of face validity
E. Potential role in decision-making		
Instructional management		
Diagnosis	Best with small group allowing detailed analysis	Must be criterion-referenced test
Placement	Acceptable with low-cost scoring	Published test often useful
Guidance	Acceptable with low-cost scoring	Published test often useful
Examinee screening		
Selection	Essential if writing is key selection criterion	Acceptable if writing is one of many criteria
Certification	Scoring scheme must reflect minimal skills	Items must sample minimal skills

COMPARISON OF ASSESSMENT METHODS

BASIS OF COMPARISON	DIRECT ASSESSMENT	INDIRECT ASSESSMENT
E. Potential role in decision-making		
Program Planning		
Survey assessment	Acceptable with low-cost scoring	Scoring efficiency an advantage
Formative evaluation	Content validity of scoring criteria crucial	Curricular validity crucial
Summative evaluation	Content validity of scoring criteria crucial	Curricular validity crucial
F. Extent of use in statewide assessments	7 states	1 state
	Combination of direct and indirect = 10 states	

NEW FRONTIERS IN WRITING ASSESSMENT

Rexford Brown
National Assessment of Educational Progress

The word frontier conjures up a number of associations for all of us, I suppose: Dodge City, Miss Kitty, the gunfight at the OK Corral. But if you live in this part of the country and hike up to the old gold and silver ghost towns or have lunch in the ruins of a miner's shack at 12,000 feet, you also associate frontier strongly with risk taking and with reaching out into the unknown. It is with this aspect of the frontier--the frontier as the edge of the unknown--that I'm concerned today, rather than the frontier as an outpost of the known. Our two previous speakers have already covered a lot of what's known. I want to call attention to what we do not know about writing and writing assessment as a way of sketching an agenda for future assessments (assuming, that is, that they are not abolished).

My remarks grow out of an AERA symposium entitled, rather timorously, "What Do We Need to Know Before We Can Assess Writing?". The researchers who pulled the symposium together were assuming that no writing assessment could take place until the results of research on the matter were all in. I was almost embarassed to tell them that the National Assessment and many

others had been assessing writing for years without waiting for research to unlock the mysteries of the human mind. But the game at the symposium was to focus on the unknown, not the known, and that is often a useful, as well as entertaining, thing to do. We who assess writing are obliged to have a viable construct of writing in mind and we are obliged to use research for deeper understanding of the process and clues about potential new approaches to its assessment. So the game is worth playing, as long as everyone understands the truth of the old saying that "the more ya know, the more ya realize you don't know."

Basic Issues Behind Our Assessments

It's important to qualify a talk such as this one--which will focus on the unknown--with several sobering observations.

The first is that we will never feel we know enough to make definitive inferences about writing ability from assessment data alone. Consider five basic questions which most of us would want to answer before we could feel truly comfortable about the construct behind our assessments:

1. How do children learn the language skills essential to writing and how to those skills develop over time?
2. What are the schools trying to teach?
3. How will we know success when we see it?
4. What psychological and social factors will affect writing performance?
5. How can we manipulate those factors in order to force explainable success or failure or error on a test?

My feeling is that even if we were able to get definitive answers, we would still have to defend our interpretations of results against attack from many quarters. The answer to the first question is going to be profoundly complex. My guess is that if the necessary unified learning theory ever appears, it will be fully understood by only a handful of

people whose expertise runs deep across such areas as sociolinguistics, psychoanalysis, anthropology, philosophy of language, measurement, cognitive psychology, literary criticism, rhetorical theory and much more. The reduction of the model to something practical will involve a number of simplifications, many of which will be controversial.

The answer to number two (what are schools trying to teach?) seems easier, but it is also very complicated. Our work in developing educational objectives persuades me that the schools are trying to teach a number of things about any subject area, some of them contradictory. And there is a difference between what the teachers think they are teaching, the school board wants taught, the taxpayers want taught, etc. The best we can hope for is a list of objectives that state what some teachers are trying to teach in some schools at some particular time; but the more specific we get, the less acceptable the objectives will be.

The third question (how will we know success when we see it?) is simply asking for a definition of good writing. We all know how elusive that is. Again, the answer is complex and highly contingent upon a multitude of factors. As for questions four and five (factors affecting successful performance on a given assessment) the combinations and permutations of psychological and experiential factors that impinge upon linguistic productions are such that I am skeptical about ever getting them under control. Indeed, I think I would worry about freedom if someone detailed and mastered them rigidly enough to determine which linguistic choices would follow which stimuli.

More knowledge does not <u>necessarily</u> help us make more valid inferences from assessment data and more useful generalizations. Indeed, too much knowledge could even make things worse. We criticize the construct and content validity of many of today's tests because our knowledge has

outstripped their premises. This will always be the case. The requirements for test design, administration, scoring and interpretation being quite rigid, we will always be forced to do a certain amount of reduction and translation of research findings into simpler and different (sometimes incompatible) forms. I have seen radical changes in theories of reading come and go while the tests stay very much the same. For instance, the National Assessment has just developed a test of reading on a psycholinguistic model. Although the distribution of items over a range of materials may be different from earlier assessments, individual items remain very much like those written from other theoretical models. The test per se has not changed greatly, but the interpretation of its data will undoubtedly reflect a psycholinguistic perspective. Research can uncover new material, theory can change, but assessment instruments may well be unaffected. This is not necessarily bad.

A second curb on anxiety about knowing enough should come from the observation that we do not need to know the etiology of cancer in order to recognize most forms of it and treat them. There are very gross indicators of writing problems just as there are tumors; both are easy to recognize and both can be treated, often quite effectively. Although we have much to learn about writing, we are neither ignorant nor helpless, even though our assessments may be unsatisfactory in one way or another.

Third, interpreters of test data should keep in mind the fact that tests provide strictly limited kinds of information. We do not have inference problems just because a particular test lacks validity; we have inference problems because it is a single summative measure. The notion that a single indicator is ever going to be a sufficient base for responsible, comprehensive inference about a complex behavior is mistaken from the start.

Interpretation of assessment results is an art, not a science. The amount of other information we bring to the task of interpreting outcome data may well influence the validity and utility of our generalizations more than anything else; indeed, it may influence what we are even able to see in the data.

So much for qualifiers. Now back to the frontier and the realms of the unknown that we will be moving into in the years ahead. The first realm is the largest of all--the culture within which writing abilities develop. Writing assessors of the future will need to know more about the cultural context of writing in order to make their most useful extrapolations from the data.

A Cultural Context for Writing

Why do AERA and NCME and Time and Newsweek care about writing today much more than they did ten years ago? Will they care in another ten years? Why is there a Right to Read but no Right to Write? What do we need to know about the politics of assessment in order to know what we need to know about writing? These are not idle questions. If we want to know what we need to know about writing abilities in order to improve assessments, we had better start by asking what we need to know about the culture that identifies this need under certain historical circumstances and defines the abilities in particular ways at particular times. Writing has had a curious history in our country, being associated at various times with penmanship, the education of the elite, Ciceronian rhetoric, grammar, good manners, good thinking, bad manners, no thinking, communication, and much more. At times it has been a frill for the college bound, at times a basic skill; not too long ago, the McLuhanites

pronounced it terminally ill. Today it seems to be getting a second wind, but Alvin Toffler has just repredicted its demise.

I'd want to know more about these ups and downs and what they mean before I invested a great deal in the future of some particular line of research, theory or assessment. We're not dealing with an ordinary subject matter here--we're dealing with a skill the secrets of which are inseparable from the secrets of the human mind itself. Ultimate knowledge about how writing works would be knowledge that far transcends mere classroom applications; it would be knowledge about how people think and how we can teach them to think. Neither a repressive society nor a society determined to make everyone happy nor Faust could ask for more.

Thinking about writing in this extreme way may help us understand why it is treated in a sort of approach-avoidance way by society. We are, as individuals and as groups, ambivalent about language itself. After all, written words start revolutions; bring Presidents down; stir loins; swell the throat with tears; bring on nausea or laughter; legitimize births; marry people or divorce them; and, all too often, open doors to the unconscious most people would rather keep closed. Language is too central to human experience, too powerful, too capable of liberating at any amount even these most securely imprisoned, not to be feared, and deeply feared. I'd like to know more about our cultural attitudes toward language--spoken or written--and how these attitudes are embedded in our institutions, change over time, and influence the ways we choose to assess writing.

Attitudes Toward Language and Writing

A second realm of the unknown includes many questions about attitudes toward language and writing. We certainly need to know more about attitudes toward both language and writing before we can make our

strongest inferences from writing samples. For instance, of the many things language is for a given individual--a means of controlling experience, a means of defense, an expressive outlet, a means of discovery, a vehicle for fantasy, a means of pleasing others, something that sounds nice, etc.--which carry more weight than others? Do some carry more weight at different ages than others? Of the many things writing is for an individual--a means of communication, an expressive outlet, a punishment for talking in class, a problem in sound/symbol relationships, an act of courage, a struggle with someone else's dialect, a schooltime activity unrelated to life, etc.--which carry more weight with particular individuals, groups or ages? What influences the weighting? How does it change?

How does a student or a group of students feel about academic writing? How do they feel about tests? About teachers of writing? How do these feelings influence the kind of writing they do in schools or on tests? The National Assessment asks students to write essays under test-like conditions, but tells them their teachers will never read the essays and they will never be given a grade for their work: how is the writing thus produced different from (a) "ordinary" writing, or (b) "ordinary" academic writing. The Assessment always qualifies its results with the observation that the writing samples were gathered under special circumstances, but we're not sure what that qualification means or even if we need to bother with it. What similar assumptions does your assessment make? We all need to know more about the consequences of gathering samples under various conditions. The assessment of the future might have a built-in research component aimed at specifying the assessment context and thus qualifying the results.

We also have to know what we have lost when we analyze essays without knowing what attitudes have influenced the writing. Any bad writing sample could be bad for many reasons. Perhaps the writer cannot write well (in what respects?). Perhaps the writer doesn't want to write (why?). Perhaps he did not want to write on the day of the test. Perhaps she is afraid of the risk writing involves. Perhaps she found the subject boring. Perhaps there was insufficient time or the topic was poorly chosen or the directions were unclear. Perhaps the writer cannot think under pressure. You can add to this list your own speculations. Don't leave out the possibility that the student can't write.

How Much Writing is Enough

A third realm of the unknown includes questions about what to base inferences about writing on. Clearly, there are always problems drawing inferences about "ability" from samples of essays. The common sense way to higher ground is to learn more about the writer's attitudes and to base inferences upon more than one essay per student. But if we're going beyond one, where do we stop? For an inference about overall writing ability, shouldn't we require at least two essays per discourse mode? And, within discourse modes, are there not sufficiently different tasks to require even more? What happens to the idea of a "test" once we start talking about the need for many different kinds of writing done under various conditions? It begins to sound like a course, not a test.

At the National Assessment, we have had frustratingly concrete experience with this dilemma. Critics of the first assessment said we could not infer much about writing because we did not ask students to do all kinds of writing. We have expanded the assessment to cover all the discourse modes and required a great many different writing tasks; now we

find we have lots to say about certain kinds of expressive writing, certain kinds of persuasive writing and so on, but no overall statement. Someone could read some of our hundreds of thousands of essays and all our reports and make an impressionistic, though probably accurate, judgment about writing ability, but that would not be an assessment result, per se. However much this upsets people who want us to churn out a single magic number, I'm afraid it's the best we can offer. Personally, I'm content with the complexity of our results and the difficulty people have in generalizing from them because I think writing is a complex matter. People who are looking for simple answers must be politely but firmly encouraged to ask more sophisticated questions.

Component Skills in Writing

A fourth area of the unknown includes questions about component skills. It is fun and sometimes important to speculate about the component skills of reading and writing, but I'm doubtful that good future assessment will depend upon more knowledge about component skills. Arguments about what can be inferred about readers from vocabulary tests, reading speed tests, comprehension "sub-skills" batteries and so on, persuade me that it is unwise to wander into this territory. Can component skills be measured independently? Do they have psychological validity? What relationships exist among them? What conditions and attitudes affect which component skills in what ways, with what clear implications for teaching? Forty years of conflicting validity studies and debates in reading should chasten anyone who believes we can generalize responsibly from one or two or a few component skills. The various sentence-combining camps, flushed with early success, must be very careful now about their next steps. It is still unclear what kind of

"skill" sentence combining is, and how it related to other aspects of composition. I would hate to see people making judgments about writing progress solely on the basis of successful performance of sentence combining feats. I have already heard of people who grade students solely on the basis of changes in syntactic maturity indices, and that gives me the willies.

Understanding Patterns of Development

Future writing assessments will undoubtedly draw upon a growing body of information about development. Current research indicates that there are some developmental patterns involved in the way children learn to spell. Both common sense and research tell us that individual preferences determine which syntactic structures people use in unfair circumstances. We also know that sentence structure choice is affected by task and mode of discourse. We have indications that children have limited discourse abilities that change in breadth and duration as they mature and some evidence that children become able to write for increasingly remote audiences about increasingly abstract subjects as they grow older. The ability to integrate ideas--to make relationships between relationships--seems to be age related, as well.

Such research findings as these have some obvious applications to teaching and testing and some less obvious dangers. All suggest the folly of basing inferences about a student's writing on too few samples, or on many samples of the same kind of writing. All suggest that we can expect considerable variation in skills in any given classroom. And all suggest that certain concepts are best introduced and tested at certain ages. This last is a potentially dangerous notion when one looks at the history of American education. Educationists have an unfortunate tendency to turn

even tentative research findings about development into curricular strictures that inhibit growth. If a researcher says "It appears that skill X did not develop in our sample group until age 12," an educationist hears "Do not teach skill X to anyone under 12 years old"--a very different message. Assessment designers can use developmental information to sharpen their questions and broaden their coverage of skills; but they should never decide to omit a question because research suggests that "most children" do not possess a particular skill.

Like component skills information, however, developmental research does not translate easily into forms test developers are most comfortable with. Much of it cannot be operationalized or turned into behavioral specifications without a trivialization that robs the results of significance.

Future Directions for Assessment

Today, there is widespread concern that an accelerated growth of knowledge in linguistics, cognitive psychology, semantics, rhetoric and related disciplines, coupled with a decline in the persuasiveness and utility of behavioristic learning models has left the testing and assessment community somewhat behind. While researchers, critics and philosophers are elaborating increasingly rich multidimensional models of language, assessment finds itself focused upon an area of "the testable" that is diminishing steadily and growing more remote from important professional concerns. The assessment of the future will have to close that gap.

What are some of the unknowns beyond the frontier and what do they suggest about the directions in which assessment might go? I've listed five areas, so far, that we are going to have to learn about:

* The cultural and political context of writing assessment as an institutional activity.

* The influence of attitudes toward language and writing upon performance.

* How best to assess a skill that is so broad, so dependent upon context and so difficult to evaluate; and, allied with that, how best to draw inferences from component skills and how to link skills together for inferences about general ability.

* How best to take advantage of developmental research without trivializing or distorting it.

* And, in general, how to bring a field with strong ties to behaviorism into an alliance with scholars, researchers and teachers who find behavioral theories and principles inadequate or inappropriate to the subject of writing.

The first suggests that writing assessors of the future will be far more sensitive about the essentially political nature of their data gathering. In addition, I suspect they will be far more sophisticated and knowledgeable about the differences between standard English and various Black and Spanish dialects of English. Essay diagnosis will take into account the influence on writing performance of bidialectal and bicultural interferences partly because it will be scientifically desirable to do so and partly because it will be politically necessary.

The second area suggests that writing assessment in the future will involve the assessment of attitudes much more than it does now. Attitudes toward written language differ from culture to culture, time to time, socioeconomic class to socioeconomic class, and age to age. As we discover more about the relationship of deep seated attitudes and apprehensions to writing performance under different circumstances, we can expect that information to bear on the interpretation of assessment results. This means we are likely to see more Likert scales, attitude inventories and apprehension indicators in the future and more couching of recommendations in the context of that kind of information.

For the third area, one doesn't need a crystal ball. The more we learn about writing, the harder it becomes to design a single assessment instrument that can embrace a sufficient amount of it to please the majority of assessment constituents. There is no such thing as writing anymore. There is pre-writing and post-writing and free-writing and rewriting and writing for particular audiences under particular constraints in particular registers and modes. There is grammar and usage and penmanship and note taking and paraphrasing and, of course, spelling. If you design an assessment to address some of these skills, you are criticized for leaving the others out. The future seems to promise only two ways out of the dilemma: either we will see increasingly comprehensive assessments that incorporate more and more aspects of writing, or we will get extremely specific assessments explicitly aimed at a few important skills and making no claims about writing in a global sense.

The component skills problem is similar to this one. Are reference reading skills, outlining, sentence combining, vocabulary, spelling and topic sentence creation component skills of writing in any scientifically defensible sense? If not, what inferences can be drawn from an assessment of any one of them or any combination? I doubt that future writing assessors will work this out. Instead, I would guess that they will always have a holistic dimension to writing assessment that will lend meaning to performance on any of the component tasks. Writing will remain communication from one human being to another and assessment will always require, at some point, a human judgment about the clarity and quality of the communication. The indirect measurement of writing skills, once the primary means of assessment, will increasingly become secondary to the kinds of judgment for which no correlations can substitute.

I expect future writing assessments to be more in tune with developmental research, assuming that the different developmental models even now available can be unified or at least made compatible with one another. Already there seems to be considerable evidence that narrative and expressive writing skills follow rough developmental sequences, implying that certain skills are probably best tested at certain ages. If current assumptions about technical constraints on testing prevail, little can be done to prevent the distortion and trivialization of this knowledge I spoke of earlier. But knowledge of development plays as great a role in test interpretation as it does in test development. At the least, assessors of the future should be better able than we are to interpret their results.

The last unknown to be faced may prove to be the greatest challenge. The people who teach writing, and the scholars and researchers in fields actively studying writing--linguists, cognitive psychologists, rhetoricians and literary critics--do not, and will not in the future, share some of the most important assumptions upon which educational testing and assessment rest. In the past, the pressure has been on the writing specialists to operationize their concepts and somehow reshape their subject matter into terms compatible with test and measurement technology. In the future, as the distance between our knowledge about writing and our current approaches to testing it widens, the pressure will reverse. Assessors will be urged to reshape their thinking about test and measurement technology into terms more compatible with the complex subject of writing.

As always, the future is happening today. Here we are, sharing our current knowledge and our speculations about what lies beyond the frontier. Through the work of people like David Cohen and Walt Haney and

some of the more thoughtful critics of testing, we are already beginning to understand the politics of assessment. From the research of people like Brozik, Emig and Petrosky we're becoming increasingly aware of the role attitudes toward language can play and the utility of attitude and apprehension scales. With its broad definition of writing skills and its increasingly thorough approach to essay description, the National Assessment is pioneering comprehensive assessment models. Developmental theories have entered assessment through James Britton and the London University approach to writing instruction. The work of Hunt, Mellon, Loban and O'Hare on the development of syntactic fluency has led to several important assessment techniques--most notably sentence combining--and essay descriptive strategies such as the one NAEP is currently employing. And finally, this meeting itself, bringing together researchers, assessors and teachers of writing, represents a step toward the synergism we will need to best meet the challenges we face and best serve our own needs and those of the youngsters we are ultimately trying to educate.

PART TWO: TRAINING WORKSHOPS

- An Introduction to the Scoring of Writing Samples
- Holistic Evaluation of Students' Writing
- Primary Trait Scoring
- Application of Generalizability Theory to Examining the Dependability of Writing Sample Ratings

INTRODUCTION

During the morning's presentations, participants were provided with some general perspectives on the processes of writing assessment; in the afternoon participants turned their attention to practical and specific training considerations. Four workshops were presented, each designed to introduce some aspect of the direct assessment of writing proficiency. They included an introduction to various writing assessment scoring methods, training in the essentials of conducting holistic scoring and primary trait scoring and procedures for estimating the reliability of writing sample ratings.

In the hour and one-half working sessions, participants reviewed the characteristics of the various scoring options and then applied the respective scoring criteria and procedures to selected samples of student papers. Since each session emphasized the actual hands-on experience of assessing writing samples according to the specified scoring methods, the following seminar summaries convey only a small portion of the information covered in the workshops. The sample training materials included in the summaries are meant to be illustrative and readers are reminded that these materials should not be used to train others without further in-depth training in the method of their choice.

AN INTRODUCTION TO THE SCORING OF WRITING SAMPLES

Presented by
Beverly Anderson and Richard Stiggins
Clearinghouse for Applied Performance Testing

This session, intended for those with limited previous experience in writing assessment, provided a succinct introduction to the various options in scoring writing samples. Beginning with a brief explanation of the characteristics of three frequently used and alternative scoring methods--holistic scoring, primary trait scoring and analytical scoring--participants then delved into the unique characteristics of each method. Workshop participants received and reviewed sample scoring guides for each of the three approaches presented and then scored a series of student papers using the respective criteria for performance.

Each of the three methods discussed was examined in respect to its unique and distinctive characteristics, the advantages and disadvantages of the approach, the length of time normally required to score papers with each approach, and the planning and training requirements called for with the scoring procedures. A major portion of the time in this workshop was devoted to the practical experience of scoring student papers. Materials distributed to participants to guide the scoring process are provided on the following pages. Readers are reminded in reviewing these materials, that they are hypothetical guidelines and should not be automatically reproduced for use in other contexts. These materials, however, offer general guidelines for establishing scoring criteria and, with modification, may be appropriate in other settings.

HOLISTIC, PRIMARY TRAIT AND ANALYTICAL SCORING GUIDES FOR ASSESSING WRITING

Holistic Scoring: Introduction

Holistic scoring involves reading a paper for an overall or "whole" impression. While specific traits such as organization, syntax and originality undoubtedly influence a reader's judgment, none of these traits is directly, individually addressed. A reader makes a judgment in much the same way that he or she decides whether a novel or an essay in Time magazine is superior, mediocre, or slipshod.

In scoring, you are asked to make no marks upon the paper. The objective is to score quickly, and to trust a first impression rather than to deliberate extensively.

As a guide, you will be presented with "range finders". These are actual student papers, carefully selected for their representativeness. With a four-point scoring scale, as in this demonstration, there will be a range finder for each of the four score levels: i.e., 4, 3, 2 and 1. These range finder papers are considered so typical of papers at each of these score levels that virtually all trained readers would agree on the score each paper would receive. This is important since these papers will serve as models to assist you and the other readers in assigning scores.

Keep in mind that range finder papers are intended to represent (approximately) the midpoint of the range. For example, a 4 range finder should not be considered the best possible paper from the sample. Rather,

it should be considered a "middle 4"; that is, some 4's in the sample will be a little better, some not quite so good.

Range finders are intended as guides. Clearly, not every contingency can be anticipated. Readers are encouraged to (1) trust their own judgment, and (2) confer with a head reader to resolve any unusual problems. For purposes of this demonstration, trust your first impression in assigning scores, even if you feel a little uncertain. You will have an opportunity to ask questions after scoring.

0 Paper: "Zero" is a special category. It is not an indicator of quality (no paper is qualitatively below the 1 level). Zero should be assigned to papers that are blank, totally illegible, written in a language other than English, or that address a completely different topic (e.g., "My Dog Fred").

Analytical Scoring: Introduction

Analytical scoring is a trait-by-trait analysis of a paper's merit. Individual traits--usually five or six--are those considered important to any piece of writing in any context. In other words, these traits would be important whether one were scoring an essay on migration of the Blue Whale or a letter of complaint to a local department store.

In this instance, students have been asked to write a letter expressing an opinion. You will score the letters on three traits: ideas, organization, and wording. These are three of the five significant traits that researcher Paul Diederich isolated through his study.* In an actual analytical scoring session, you would probably be asked to deal with five or six traits.

---

* By soliciting feedback from experienced scorers trained in holistic procedures, Diederich isolated five characteristics which scorers seem to focus on in rating papers: ideas, mechanics, organization, wording and flavor. See Spandel, V. and Stiggins, R.J. Direct measures of writing skill: Issues and applications. Portland, Oregon: Northwest Regional Educational Laboratory, 1980.

The traits should be scored one at a time. In scoring, take care not to let your impression of one trait (say mechanics) influence your scoring of another trait (say organization). The following pages present guidelines for scoring each of the three traits you are asked to deal with. Remember that the definitions presented in this guide (like those in the Primary Trait Scoring Guide you will see later) were developed specifically for use with this test exercise. While the factors to be scored (i.e., ideas, organization and wording) have universal application, it would be necessary to review and (slightly) revise the guide for use with a different test exercise.

Readers are also cautioned to use these guidelines with flexibility. It is not necessary that a given paper conform totally to the definitions presented for a given score level. Every paper is different; not even the best guidelines can account for every variable. Therefore, assign scores according to which definition seems best to fit a given paper.

Factor 1: Ideas

Definition: For purposes of this test item, "ideas" may be defined as arguments or opinions, or any thoughts expressed in support of those arguments/opinions.

Criteria:

4 Paper: A 4 paper offers sound ideas that indicate the writer has given careful consideration to both sides of the argument. The stand taken may be strongly expressed, but it is clearly the result of careful thinking rather than mere impulse. Occasionally, the ideas expressed will be highly original, or will be presented in an original, insightful, or entertaining manner.

3 Paper: A 3 paper resembles a 4; the ideas are simply not as sound or as striking. There is some evidence the writer has given thoughtful consideration to his/her arguments, but the paper as a whole is likely to be less insightful or persuasive, and more emotional. The ideas are clearly presented, but there is little or no evidence of originality (e.g., humor, unexpected insight).

2 Paper: A 2 paper offers little or no evidence of careful judgment or insight. Ideas may seem arbitrary or whimsical, or may be expressed in strong emotional terms (e.g., "I hate this idea!"). Clarity may be a problem too. Often the paper will offer little more than one idea restated repeatedly, or the writer will fail to make his/her stand sufficiently clear. Originality is virtually absent; ideas may be little more than opinions attributed to others (e.g., "My parents think..." or "The teacher won't like it if..."). There is no effort to persuade the reader through the choice of ideas or effectiveness of presentation.

1 Paper: In a 1 paper, clarity may be such a problem that the reader can scarcely discern the writer's viewpoint. Often the ideas expressed are nothing more than a restatement of the question. Sometimes, the writer will be highly ambiguous or will appear to take two sides at once. There is no evidence that the writer has given real thought to the presentation. Ideas are generally arbitrary or do not stand out clearly. The writer's statements are generally ambiguous, muddled, or indecipherable.

Factor 2: Organization

Definition: For purposes of this test item, "organization" may be defined as a sense of order, ability to stay on topic, and ability to relate all details to a central idea or argument.

Criteria:

4 Paper: The 4 paper is carefully organized from beginning to end. Main points are clearly expressed with no apparent contradiction or shifts in viewpoint. The writer takes a clear stand and supports it with relevant evidence. Details are presented in logical order. Important details are not omitted, nor does the writer wander from the topic.

3 Paper: The 3 paper is generally well organized, though there may be some lapses. Main points can be readily inferred but may be less clearly expressed than in a 4 paper. The writer takes a definite stand for or against the 12-month school year; however, supporting evidence may be weak, irrelevant or disordered in some cases (e.g., "I forgot to mention that..."). The writer may wander from the main point once or twice, but the reader can readily follow the argument presented.

2 Paper: The 2 paper shows some attempt at organization, but the result is ineffective. Details are often presented in haphazard or random order. Shifts in viewpoint are

common (e.g., the writer appears at first to favor the 12-month school year, but winds up opposed). Some writers never take a clear stand pro or con. Supporting evidence--if any--tends to be sketchy or disjointed. Important details are often omitted. The writer may have trouble staying on topic. A 2 paper may be little more than a fragmented list of details that are not clearly related to a central theme or to one another.

1 Paper: The 1 paper may be so short that it defies organization, or so unclear (muddled, ambiguous) that the reader can discern no reasonable pattern of thought. Often, the reader cannot tell whether the writer favors or opposes the 12-month school year. Any "evidence" presented may be little more than a restatement of the writer's opinion (e.g., "I am against the 12-month school year because it isn't a good idea"). In a paper of any length, the writer will tend to wander from the topic--often bringing up a whole new subject. In short, the 1 paper is too brief or disjointed to allow any real organizational structure.

Factor 3: Wording

Definition: The purposes of this test item, "wording" may be defined as vocabulary and word choice.

Criteria:

4 Paper: A 4 paper exhibits strong, carefully chosen vocabulary. Words are correctly used and often selected to convey a precise meaning or to create a particular impression or mood. The vocabulary seems to come naturally to the writer; that is the writer is not straining to use words that will impress the reader. The word choice is appropriate to the topic. There is notable absence of cliches, jargon or slang. The language is fresh, clear, effectivce--possibly persuasive.

3 Paper: A 3 paper exhibits acceptable, comprehensive vocabulary. The word choice is less striking, dramatic or effective than in a 4 paper; nevertheless, the reader is never reduced to guessing at the writer's meaning. Usage is generally correct; there may be an occasional exception where the writer attempts a construction or uses language beyond his/her level of understanding, but such exceptions do not seriously impair the readability of the paper. The word choice is appropriate to the topic; the writer may use some slang, but does not rely on slang, jargon, or cliches. While errors are rare, language is less likely to invoke a particular mood or create a striking image than in a 4 paper.

2 Paper: A 2 paper often contains numerous errors in usage and word choice. Usually, the reader is able to decipher meanings, but the errors may be frequent or severe enough to seriously impair readability. A 2 paper may also exhibit such simplistic vocabulary that it is difficult for the writer to convey the ideas he/she wishes to express. Word choice may be reasonably accurate, but inappropriate to the level of sophistication demanded in a strong argumentative paper. Use of slang, jargon or cliches is common. It is clear that the writer does not have enough vocabulary at his/her command to select effective or precise wording. A 2 paper lacks the spontaneity and flow of even a 3 paper, and can often be read only with great difficulty.

1 Paper: A 1 paper is nearly impossible to decipher. Word choice, usage and syntax are so poor that the reader can only guess at the writer's meaning. The writer appears not to have even a rudimentary vocabulary at his/her command. Often the response is little more than a restatement of the question, and may suggest that the writer did not understand what was required. Whereas the 2 paper can be understood with effort, the 1 paper often defies interpretation. The frequency and seriousness of errors may antagonize or simply baffle the reader.

0 Paper: "Zero" is a special category. It is not an indicator of quality (no paper is qualitatively below the 1 level). Zero should be assigned to papers that are blank, totally illegible, written in a language other than English, or that address a completely different topic (e.g., "My Dog Fred").

Primary Trait Scoring: Introduction

Different traits are important to different pieces of writing. For example, in a set of instructions for assembling a bicycle, the most important trait may be clarity, for unless the instructions are clear and comprehensible, it is unlikely that the reader will be able to perform the task.

For this test item, the writer is asked to write a letter to the Chairman of the Board, conveying his/her views on the proposed 12-month school year. Many things--word choice, neatness, tone, sentence structure, ideas--will undoubtedly affect the intended reader's (that is,

the "Chairman's") response to the letter, and your response as well. Perhaps the most important factor will be persuasiveness--that is, to what extent the letter is capable of persuading the reader to adopt a certain point of view.

Taking persuasiveness, then, as the primary trait, here is a guide for scoring papers on a 4-point scale.

4 Paper: The writer takes a strong stand either pro or con and defends it with solid evidence. The arguments are logical, and seemingly founded on facts or on real experiences. Tone and word choice are appropriate. Often the paper will exhibit rhetorical excellence through exceptional vocabulary, highly sophisticated sentence structure or a particularly well constructed argument. The writer seems very aware of audience. The strength of a 4 paper may lie in its style or in its logic; either way, it can truly be labeled "persuasive".

NOTE: The reader may agree or disagree with the writer's point of view; this is irrelevant and must not be allowed to affect scoring. A reader may, for example, assign a paper a 4 on the basis that--"This was a well written letter conveying a sound and highly persuasive argument--even though it does not express the position I would have taken."

3 Paper: The writer's viewpoint is clear, but supporting evidence may be less effectively presented than in a 4 paper. A 3 paper exhibits no major flaws; that is the tone and wording are appropriate to the subject and audience. However, the paper is less rhetorically effective than a 4. It will often have some identifiable weaknesses--perhaps in word choice, style, completeness or organization--but will not be weak in many areas nor seriously weak in any. Overall, a 3 paper merits consideration, but is generally not thought provoking or powerfully moving.

2 Paper: A 2 paper overall lacks the power to persuade. The reader can usually determine the writer's viewpoint, but may not be clearly expressed. Often the supporting evidence offered is weak, irrelevant or incomplete. It may be difficult to perceive any real awareness of audience; tone or word choice may seem at times inappropriate. Some 2 papers may be little more than an outpouring of emotion, with little or no effort to persuade the reader logically or stylistically.

1 Paper: Unlike a 2 paper, a 1 paper may be virtually incomprehensible or may exhibit no effort to persuade the reader. It is generally difficult to determine the writer's point of view; there is no clear expression of opinion, or whatever

statement is offered is ambiguous. Usually, there is no supporting evidence in a 1 paper, or what is offered is too limited or poorly expressed to be of value. Like a 2 paper, a 1 paper may be little more than an outpouring of emotion, but generally it will lack even the cogency evident at the 2 level. The 1 paper exhibits weaknesses in vocabulary, rhetoric, organization, logic and word choice--in short, there are no redeeming "persuasive" factors.

0 Paper: "Zero" is a special category. It is <u>not</u> an indicator of quality (no paper is qualitatively <u>below</u> the 1 level). Zero should be assigned to papers that are blank, <u>totally</u> illegible, written in a language other than English, or that address a completely different topic (e.g., "My Dog Fred").

HOLISTIC EVALUATION OF STUDENTS' WRITING

Presented by:
J. Evans Alloway
Educational Testing Service

In this workshop, intended for those desiring more specific information on holistic scoring, participants experienced a simulated training session that readers receive as they begin an actual scoring session. Individuals in the session were introduced to the rationale that supports holistic and comparative evaluation, considered the responsibilities facing the evaluator in writing assessment and some of the ever present realities of evaluation; e.g., issues of time, money, accuracy and usefulness.

Students' papers served as the medium of instruction. Using the holistic method, which requires readers to judge the writing as a coherent whole, workshop participants learned to assess the quality of a piece of writing and focus on broad aspects of writing performance. Individuals had an opportunity to judge the overall quality of numerous student writing samples. In doing so, they considered the quality of thought, organization, punctuation, sentence structure and other aspects of writing at the same time rather than separating the various aspects and judging each individually.

The general informational materials on holistic scoring, made available at the session, are included on the following page.

HOLISTIC EVALUATION OF STUDENTS' WRITING

I. What is holistic scoring?

A. The theory

1. The whole of a piece of writing is greater than any of its parts.

2. English teachers, though they may have difficulty in giving a verbal description of writing ability that is recognizable to all, can recognize good writing when they see it.

3. Though in an analytic reading teachers may not agree on the weight to be given a particular trait, these same teachers will, in judging a work as a whole, rank papers in much the same way.

4. No aspect of writing skill can really be judged independently; the halo effect is always strong.

B. The method

1. The standards

a. Standards are not imposed upon readers; readers themselves determine standards.

b. Papers are not judged against an ideal, but against what is: what students have written on this topic at this time.

c. Standards must be maintained and reinforced throughout the reading.

2. The judgments

a. Judgments are made on anonymous papers.

b. Judgments are independent.

c. Multiple judgments on each paper are mandatory.

d. Judgments must be quick and immediate.

e. Judgments must be definite, for the score scale has no middle points.

3. The scoring

a. The score is the sum of all the readers' judgments.

b. Some discrepancies in the scores the readers give are to be expected.

c. Wide discrepancies between readers' scores must be corrected immediately.

d. Regular divergence from the standards on the part of any reader must be corrected.

II. Why use holistic scoring?

A. It is efficient.

B. It is reliable.

C. It emphasizes what is right rather than what is wrong with a piece of writing.

D. It requires consensus among readers.

E. It encourages evaluation of the program, as well as the individual pieces of writing.

III. How is a topic scored? (Actual reading)

A. The topic is read and analyzed.

B. The ground rules are established.

C. The standards are set through the use of sample papers.

D. The papers are read.

1. First reader's scores must remain unknown to other readers.

2. All papers should be read once before any are read twice.

3. Readers must be allowed to rest regularly.

4. Papers must flow efficiently from reader to reader.

IV. What makes a good topic?

A. The interest to the students.

B. The interest to the readers.

C. The range of writing it produces.

D. The relative objectivity with which it can be scored.

V. Of what use is holistic scoring in the schools?

A. It can promote communication about the teaching of writing among faculty members.

B. It can be used to measure growth in students' writing ability.

C. It enables teachers to score writing assignments quickly and reliably.

D. It calls for multiple evaluations.

Some Considerations in Judging the Worth of an Essay Topic

1. Do the candidates seem interested in the topic?
   Do they all have something to say in response to it?
   Can they find specific examples to use?

2. Are the responses unduly emotional or patriotic or religious or in some other way difficult for readers to judge fairly?
   Can readers make judgments easily between essays at different levels of competence?
   Are all the essays alike--mediocre?
   Does the topic encourage cliche-ridden or "canned" responses?
   Does the topic invite candidates to try to guess at what attitude the readers want them to take toward the topic?

3. Is the topic confusing in some way?
   Are the requirements of the topic ambiguous?
   Is the topic subject to several accurate interpretations?
   Are candidates misreading or misconstruing some word or phrase in the topic?
   Are candidates following the requirements of the topic?
   Is the mode of writing the candidates use in keeping with the mode the topic is meant to produce?

4. How will readers react to the topic?
   Are the responses just plain boring?
   Can judgments be made quickly on reasonable grounds?
   Are high scores being awarded for writing ability and not for extraneous matters like sophisticated literary allusions or unique examples?

5. How long are the essays?
   Has too much time been allowed or too little?

6. Do the essays lack a starting point?
   Do the essays lack organization?
   Is the topic at fault for failing to suggest a method of organization?

7. Would you be willing to read essays on this topic for several days in succession?

Suggestions for Writing Essay Questions

1. The question should be clear.

   Students should not have to puzzle over the instructions. The topic is intended to test the ability to write the answer and not the ability to guess what the test maker intends. Besides, students have only a limited amount of time, time that should be spent writing and not analyzing unnecessarily.

2. The question should be as brief as clarity allows.

   Restatement may sometimes be necessary to avoid misunderstanding. But, then, perhaps one should consider whether the restatement should be used without the original because the restatement does not need additional clarification.

3. The instructions should be definite.

   Students should know what is required. For example, discuss, citing specific examples from one novel. Or, pay attention to the correct form of the business letter. Or, be sure to use complete sentences.

4. Avoid questions requiring only a yes or no answer.

   Example: Do you agree? Where does the student go from there?

5. Average students should be able to write average answers to the question, and yet bright students should be able to show their brightness.

   A good topic permits the ranking of all students according to ability.

6. The vocabulary used and the concepts expressed in the topic should not be difficult for the ordinary student to understand immediately.

   A difficult topic distinguishes only between the very bright and the rest of the population. Besides, difficult reading changes the test to a reading test.

7. The question should not call for cliches as answers.

   A topic worn out by overuse produces worn-out responses. On the other hand, some good questions merely twist cliches. For example: What's right with television? In what ways are teenagers more conservative than the over-thirties?

8. The question itself should provide an organizing principle for the essay.

Example: Compare and contrast...
Briefly describe...and then analyze...
Discuss your answer to this question, giving the reasons for your answer and citing specific examples to support those reasons.

9. The question should not elicit responses which affect either the writer's or the reader's judgment.

   Politics, racial issues, and other inflammatory topics are to be avoided. Also to be avoided are topics that are dishearteningly dull. For example, even a seemingly innocuous topic like: Who has had the greatest influence on your life? cannot be scored easily. If the candidate writes on the wrong political figure (from the reader's point of view), the score is either too high--because the reader is making up for his or her own bias--or too low because the reader has succumbed to that bias. On the other hand, the fifth essay on the greatness of the basketball coach is not scored on the same standard as the first. Readers are human; they do become bored.

10. The question writer should write out the answer expected and determine whether the question really calls for that answer. The question writer should also try to answer the question in the allotted time, just to see whether it is humanly possible to do so. The question should be revised in the light of any discoveries made.

PRIMARY TRAIT SCORING

Presented by:
Ina Mullis
National Assessment of Educational Progress

This session introduced participants to the process of the primary trait system and the procedures used in scoring the National Assessment's writing samples. In this approach to assessing writing, the underlying rationale is that writing is done specifically for an audience and can be judged in view of its effects upon that audience. The workshop activities focused on the background and rationale for developing primary trait scoring, its recent use in conducting the 1978-79 NAEP assessment, and its relationship to NAEP's other scoring procedures--cohesion, syntax and mechanics. Participants in the session were introduced to the procedures used to train readers, to conduct scoring sessions, and the activities involved in developing writing exercises and scoring guides for using the primary trait method. In addition, individuals had an opportunity to review some of the recent examples of NAEP's writing assessment tasks and to practice scoring student papers with the primary trait scoring guides.

The following description of primary trait scoring is an abbreviated version of a larger paper provided to workshop participants. An additional paper, Guidelines for describing three aspects of writing: Syntax, cohesion, and mechanics, is also available from Ina Mullis through NAEP.

USING THE PRIMARY TRAIT SYSTEM FOR EVALUATING WRITING[1]

The Primary Trait System (PTS) for scoring writing tasks was developed by the National Assessment of Educational Progress (NAEP) and used to rate papers from the second (1973-74) and the third (1978-79) national assessment of writing. The purpose of this paper is to present the rationale and procedures used to implement the Primary Trait System for evaluating writing.

Defining Primary Trait Scoring

The rationale underlying primary trait scoring is that writing is done in terms of an audience and can be judged in view of its effects upon that audience. Particular writing tasks require particular approaches if they are to be successful. The approach used by the writer to reach and affect his audience will be the most important--the primary--trait of a piece of writing. For example, the writer of a set of directions must present things in a logical and unambiguous manner if he expects readers to follow the directions. Therefore, the primary trait of a written set of directions would be an unambiguous, sequential and logical progression of instructions. Successful papers will have that trait, unsuccessful papers will not, regardless of how clever or well written they may be in other respects. On the other hand, the purpose of campaign literature is to persuade a reader to vote for a candidate. A successful campaign paper will have certain persuasive traits that an unsuccessful one will not

---

1 This paper is adapted from a more comprehensive report by Ina Mullis titled Using the Primary Trait System for Evaluating Writing. NAEP, 1980.

have, and these traits will differ from those necessary for a successful set of directions.

A carefully constructed testing situation provides opportunities for respondents to demonstrate their ability to choose and effectively carry out appropriate rhetorical strategies, but the effects of the strategies cannot be measured by actual success. For example, it is not feasible to have respondents write real job applications and then check to see if they were hired. An evaluation of the writing has to be based on its likelihood of achieving the desired effect. The features that would contribute to this success must be identified and defined in terms of their importance. These definitions then become the scoring criteria.

Identifying Important Writing Skills

There are many factors that can be taken into account when evaluating a piece of writing. Responses can be scored for content, organization, syntax, mechanics, originality of expression and so forth. In addition, there is a broad range of writing tasks that can be assigned. Students can be asked to write organized reports for their teachers, to engage in imaginative writing, to write descriptive letters to friends or to write basic explanatory letters, such as those required to straighten out a billing error or order something through the mail. The responses elicited by a given assignment may or may not be suitable for a particular kind of evaluation.

With the Primary Trait System, the first task is to determine which writing skills are the important ones to consider for the purpose at hand. Evaluation committees, curriculum committees, teachers and parents must decide:

1. To whom, and about what, students should be able to write.
2. What kinds of information should be provided by the evaluation.
3. What decisions can be made on the basis of the results.

It is very important that purposes for writing and the skills necessary to achieve those goals be identified and used as guidelines in developing the specific tasks that will be evaluated. Otherwise, the tasks developed may not be the best way or even be an appropriate way to measure the skills of interest. Also, the tasks may inadvertently overemphasize or underemphasize certain areas.

Developing and Selecting Writing Tasks

Once the purpose of the writing and the situation for it have been established (why write and to whom), the groundwork is in place to begin the arduous business of developing specific writing tasks. In developing any writing task, the subject matter, the audience and nature of the writer are three fundamental variables. Since the best writing stimulus is one that induces the greatest number of people to write and gives the most people a chance to write well, writing tasks often suggest a variety of subject matters or a variety of audiences in an effort to strike a responsive chord with a student. This approach to writing task development appears to make sense, yet evaluating the responses to such writing tasks can be very difficult. A multi-focused stimulus often evokes responses too dissimilar to be considered on the basis of common criteria. This creates the possibility that the internal criteria of the reader may interfere. An imaginative or inspirational paper may get a higher score than an equally well written paper using a sarcastic tone. If imagination, inspirational ideas or tone were important, the writer should have been given a task that required these traits for a successful

response, and the scorer should have been reading for them on an explicit basis.

The essence of the Primary Trait System is to narrowly delineate the situation for the writer by defining the variables. This means that for each exercise three things must be specified:

1. The identity of the writer (whether the respondent is himself or is given a role to play).
2. The audience (who the writer is writing to).
3. The subject matter (what the writer should communicate to the audience).

With this approach, responses should all address the same task and can be judged using the same criteria. Consequently, systematic judgments can provide both descriptive and comparative information about how well individuals or groups of people can use writing to communicate in given situations.

The more structured the task, the less difficult the evaluation, since the essays or letters will be more uniform in focus. For example, if the ability to persuade is chosen as an important skill to measure, it might be decided that students should be able to use writing to influence decision makers. A possible task could be--"Write a letter to someone important about a problem in this country." A better task would be--"Write a letter to your principal suggesting a way to solve a problem in your school." If the nature of the problem were even further defined, for example, the food quality in the lunchroom, the task would be even better. With a national sample, it is difficult to identify universally applicable situations. For a classroom, a school or even a school district, the problem of respondents having enough common knowledge and experiences on which to base the tasks should be less severe.

It was stated earlier that one identifies the writing skills that are desirable to measure before developing items. On the other hand, any item contrived to require the utilization of a specific skill or skills will not suffice. In defining a task in terms of a purpose, the audience and the situation, the overriding goal of any writing exercise must not be forgotten. The exercise must stimulate students to write--and further, to write as well as they can. Thus, in addition to the above features each exercise must have some motivating feature. Each task must be very carefully developed to be realistic and interesting, as well as to be a specific measure of the skill.

NAEP considers the development of a writing task to be more than just an item. For example, using the development plan, a purpose of writing is selected (persuasion for personal/social reasons) and a task is developed to measure ability to write for that purpose (to convince the landlord to let you keep your puppy). The basic trait of a successful response to the task is identified (making appropriate appeals) and then a rationale is written describing how the task relates to the purpose and the primary trait. A description outlining four levels of success on the trait is also developed. A complete PTS writing task consists of a purpose, a primary trait, a rationale, a task and a scoring guide. Sometimes secondary traits are also identified.

Next, all items are field tested in a variety of schools representing different types of programs and socioeconomic levels. When an item is judged to be successful, the scoring outline is refined on the basis of the actual responses. For example, the task developers may have had unrealistic expectations about the types of appeals that would sound most convincing. Then the task, its new scoring guide and the responses are

set aside as a unit to be considered when the time comes for final selection.

Developing Scoring Guides

Each writing task developed needs a scoring guide specifically tailored to the primary trait being measured by the exercise. Generally, four levels of competency are defined. The first and lowest category is for respondents who show little or no evidence of the skill being measured. The second category is defined as minimal evidence of the skill. Category "3" is defined as demonstrating competence and category "4" is reserved for those papers which are excellent or outstanding.

As explained earlier, once an important skill is identified and an item is developed to measure that skill, it is necessary to outline what consitutes achievement of that skill, given the specific situation of the writing task. The following are some basic outlines for primary trait score categories for the three basic purposes of writing assessed by NAEP.

In broad terms, presentation of compelling evidence and arguments is the primary trait for persuasive writing. Generally, then, a "1" paper would present no reasonable arguments, a "2" would have one reason or perhaps several not particularly thought out reasons, a "3" would be a well thought out presentation containing several appropriate reasons, a "4" paper would be well organized and support reasons with compelling details. The quality and appropriateness of the reasons depend upon the specific task. Sentimental appeals may be convincing to grandparents, but probably are not as appropriate for a principal or prospective employer. For these audiences it is likely that a more rational approach will increase chances of success. The arguments might present a type of cost/benefit analysis, repudiate the risks involved or appeal to

authorities such as history or testbooks. The specific category definitions should consider the variety, appropriateness, organization and execution of the arguments.

Selecting and Training Readers

National Assessment requires that the Primary Trait Scoring be done by readers with previous English experience. This does not necessarily mean that the scoring should be contracted out or that the readers must be English teachers. Several districts have conducted successful assessments using teacher's aides, substitute teachers, local graduate students or parents as readers. It should be noted that there are side benefits to using teachers as readers. If teachers are used, the Primary Trait training can be an in-service activity, and the teachers tend to feel less threatened and more interested in the results. Teachers may also feel more enthusiastic about working toward curricular changes based on assessment results they helped produce.

To determine the number of readers necessary, NAEP first decides how many readings are needed per paper and then compares the number of hours per day's worth of work with the assessment schedules. NAEP estimates that one reading for primary trait takes about one minute for a paper written by a 9-year-old, one and one-half minutes for a paper by a 13-year-old, and two minutes for a paper by a 17-year-old. (These time estimates include breaks and time for organizing the papers.) Of course, actual times will vary depending on the difficulty of each specific writing task, the propensity of the respondents to write and the number of secondary traits to be scored.

Any time a group of readers is used, a chief reader or group leader should be designated. During the actual scoring the group leader is

responsible for answering questions about scoring procedures and reconciling discrepant scores.

The amount of time devoted to preparing for training is the key to a successful training session. It is during this stage that the guide, prepared on the basis of field test information, is refined on the basis of actual assessment responses. This is the time that the group leader becomes totally familiar with the guide and how it is to be applied to "live papers".

To prepare for training, the analyst responsible for scoring, consultants, the scoring directory, and the group leader first review the guide and the categories assigned to the tryout papers. Next, the trainers start applying the PTS guide to actual assessment papers, independently scoring five to ten of the same papers at a time. Discrepancies are discussed and resolved, and when necessary the guide is modified. Sometimes an unanticipated type of paper appears in the assessment and needs to be accounted for in the guide. On other occasions, the original category descriptions are found to be unclear, vague or ambiguous. This process continues, usually through fifty to one hundred papers, until the group is in agreement, the guide is refined and enough papers have been prescored to prepare a training packet.

The process of achieving a detailed understanding of the guide is most important. The actual scoring must be done with clear reasoning directly based on the philosophy underlying every scoring category. This does not mean that every paper will be a clear cut example of a particular category. But the trainer(s), and subsequently the readers, must have a firm enough foundation to not only recognize features of "2" papers in "2" papers and features of "3" papers in "3" papers, but to recognize "2" and "3" features in "2/3" papers. They must be able to reason and explain why

such a paper should be finally designated as either a "2" or a "3". NAEP has found it very helpful to take notes on specific papers and use them during training. In this way, readers not only have the guide and practice in reading papers, but they have a record of the reasoning behind the classification of particular papers.

Scoring the Writing Tasks and Reliability of the System

In scoring actual assessment data, each paper is scored by two readers and discrepancies are reconciled by a third. The need for reconciliation usually occurs for less than 10 percent of the papers.

The chief reader usually spends time in the early scoring reading papers already rated by each of the readers. In this way, misconceptions can be detected and clarified early in the scoring session. Also, assigning the chief reader responsibility for reconciling disagreements is another good way to determine whether or not particular readers have misunderstood directions.

By using the methods for developing PTS exercises and scoring guides and the procedures for training readers described in earlier sections of this document, National Assessment maintains percentages of agreement between first reader and second reader that are over 90 percent. For example, for the 1978-79 scoring effort (21 essay sets, each with 2,500 responses) percentage of agreement ranged from 91 percent to 97 percent with an average percentage of agreement of 94 percent. These consistently high percentages of agreement suggest that for purposes of estimating the performance of groups of persons at one point in time, for example, the eighth graders as compared to the fifth graders, one reading might be considered sufficient.

Of course, for any essay evaluation two readers are still recommended. The results of the two readings can either be summed, summed and averaged, or reconciled by a third reader as in the process used by National Assessment. In cases where the purpose is to estimate performance of individuals, not only should at least two readers be used, but as many writing samples as possible across a variety of tasks should be evaluated for each individual.

APPLICATION OF GENERALIZABILITY THEORY IN EXAMINING THE DEPENDABILITY OF WRITING SAMPLE RATINGS

Presented by:
Robert L. Brennan
American College Testing Program

The objective of this session was to instruct participants in some relatively straightforward applications of generalizability theory to issues that arise in determining the dependability of writing sample ratings. As reviewed in the seminar, generalizability theory employs designs and procedures that are similar (although not identical) to those discussed in traditional treatments of analysis of variance. Specifically, for writing samples, data are sometimes collected according to a design in which each rater rates all writing samples (or all aspects of a writing sample) for each examinee. In an alternative design, data are sometimes collected according to a process in which each rater rates only some writing samples (or some aspects of a writing sample) for each examinee.

In considering these two designs, special consideration was given to estimating and interpreting quantities such as variance components, error variances, universe (or "true") score variances, and generalizability coefficients (which are reliability-like indices). Participants also had an opportunity to perform the analysis using sample data sets.

The following paper provides a brief introduction to generalizability theory including consideration of most of the issues discussed in this session:

Brennan, Robert L. and Michael T. Kane. Generalizability theory: A review. New Directions for Testing and Measurement, 4, 1979, 33-51.

APPLICATION OF GENERALIZABILITY THEORY TO EXAMINING THE DEPENDABILITY OF WRITING SAMPLES RATINGS

Summary of Training Session No. 4
Instructor: Robert L. Brennan, ACT

The objective of this session was to instruct participants in applications of generalizability theory to issues that arise in determining the dependability of writing sample ratings. Generalizability theory is an approach to examining the consistency or dependability of educational measurements. If differs from classical test theory in that, whereas classical theory says that a person's observed score is a combination of a true or universe score and a single error compenent, in generalizability theory an observed score is a composite of a universe score and (possibly) several error components. Thus, using classical test theory, different types of analyses are necessary when one is interested in different types of reliability (stability, equivalence, inter-rater, etc.). In such cases, the single error component in each reliability formula pertains to a different source of error (across time, across test forms, across raters, etc.). Because generalizability theory allows for multiple error components, issues associated with these different types of reliability can be addressed concurrently.

As reviewed in the training session, generalizability theory employs designs and procedures that are similar (although not identical) to those discussed in traditional treatments of analysis of variance.

In generalizability theory, however, emphasis is placed upon estimating variances (called variance components) associated with each "main" effect and "interaction" effect in a design. Then, the theory tells the investigator how to combine these variance components to obtain estimates of universe score variance, different types of error variance,

and generalizability coefficients (which are reliability-like indices). It is particularly important to note that this approach makes it possible for an investigator to quantify how different potential sources of error (e.g., raters and/or writing samples) contribute to error variance.

The implications for writing assessment are obvious. If one can separate the error associated with both items (writing samples) and raters, persons planning the assessment can modify their procedures to produce the smallest possible error in the scores, given time and financial resources available. In the training session, two specific scoring designs were described. One design occurs when each rater rates all writing samples (or all aspects of a writing sample) for each examinee. In the other design, each rater rates only some writing samples (or some aspects of a writing sample) for each examinee. In notation presented in the session, these two designs are identified as $\underline{p} \times \underline{i} \times \underline{t}$ (the first design) and $\underline{p} \times (\underline{i} : \underline{t})$, where $\underline{p}$ represents examinees, $\underline{i}$ represents writing samples, and $\underline{t}$ represents raters.

In considering these two designs, sample data sets were examined to illustrate changes which occur in error variances and generalizability coefficients ("reliability") when the number of writing samples an examinee provides and/or the number of raters scoring each writing sample, are modified: In general, as the number of writing samples and raters scoring each sample increase, generalizability co-efficients increase and error variances decrease.

But time and money costs are real parameters within which designers of writing assessments must work, and there is a limit to the number of samples and raters one can afford. Persons designing writing assessments can study the relationship between number of writing samples and number of raters and dependability to decide how many samples each examinee should

provide and how many raters should score each sample to achieve a desired level of dependability. If, for a particular assessment, it is desired that an estimated generalizability coefficient of at least .70 be achieved, assessment planners can decide whether it is less expensive to have each examinee provide six writing samples, each of which will be scored by two raters; or to provide four writing samples, each of which will be scored by four raters.

In summary, generalizability theory was shown to be quite applicable to issues associated with writing assessment. The theory's pay-off lies in the fact that error associated with different sources (such as writing samples and raters) can be differentiated rather than being combined into a single error component as in classical test theory. Once the effect of the different sources is known, procedures can be varied systematically until the optimum combination (in terms of both measurement qualities and practical constraints) of number of samples and number of raters can be identified. The end result will be increased accuracy of judgments about a person's "true" writing ability based on his/her observed writing assessment score.

APPENDIX

List of Seminar Participants

Vern Achtermann
Assessment Services Coordinator
NAEP
1860 Lincoln Street, Suite 700
Denver, CO 80295

Russ Allen
Department of Public Instruction
126 Langdon Street
Madison, WI 53702

Robert L. Arnold
Educational Psychology
Vought Corporation
1501 Morning Drive
Loveland, CO 80537

Marilyn Averill
Evaluation Specialist
Boulder Valley Schools
P.O. Box 9011
Boulder, CO 80302

George Barnes
Evaluation Consultant
Nevada Department of Education
400 West King
Capitol Complex
Carson City, NV 89710

Merle B. Bass
Educational Consultant
Jefferson County Schools
2631 S. Brentwood Court
Lakewood, CO 80227

Barry B. Beal
Supervisor
Dept. Development & Evaluation
Denver Public Schools
900 Grant Street
Denver, CO 80203

Roscoe Boyer
Bureau of School Services
School of Education
University of Mississippi
University, MI 38677

Richard Brengle
805 S. First Street
Ann Arbor, MI 48103

Donna Brennan
Director of Curriculum
Cherry Creek Schools
4700 South Yosemite
Englewood, CO 80111

Mervin M. Brennan
Assessment Specialist
Illinois State Board of Education
100 North First Street
Springfield, Il 62777

Catherine Brigl
Education Diagnosis
Jefferson County Public Schools
10969 W. 62nd Place
Arvada, CO 80004

Evelyn Brzezinski
Assessment & Measurement Program
NWREL
710 S.W. Second Avenue
Portland, OR 97204

Jerald M. Bullock
Educational Program/Policy Analyst
Georgia Educ. Improvement Council
Room 656
7 M.L. King, Jr., Drive, S.W.
Atlanta, GA 30334

Kathleen Burk
Associate Director
Regents' Testing Program
Box 868
Georgia State University
Atlanta, GA 30303

Mary Callahan
National Evaluation Service
P.O. Box 226
Amherst, MA 01002

Patricia T. Candal
Asst. Prof. of English Education
University of South Carolina
Coastal Carolina College
Conway, SC 29526

Jane Christensen
Chairman
Language Arts, Resource
Jefferson County Schools R-1
5916 Urban Street
Arvada, CO 80004

Rebecca S. Christian
Asst. Director
Bureau of Accountability
Louisiana Dept. of Education
P.O. Box 44064
Baton Rouge, LA 70804

Richard N. Claus
Statistical Analyst
Saginaw Public Schools
3207 Curtis Road
Birch Run, MI 48415

Barbara Cole
Coordinator
Research & Assessment
Oregon Department of Education
700 Pringle Parkway, S.E.
Salem, OR 97310

Mary Cooper
Research Officer
B. C. Research
3650 Wesbrook Mall
Vancouver, B.C., Canada
V6S 2L2

Constance S. Dardin
Coordinator
Student Assessment
Department of Education
Little Rock, AR 72201

Elizabeth Davis
2227 Canyon Blvd., #102
Boulder, CO 80302

Constance C. Dolese
Director of Secondary Education
New Orleans Public Schools
4100 Touro Street
New Orleans, LA 70122

Sandra Edleman
1517 S. Theressa
St. Louis, MI 63104

Christine Ewy
Consultant
Second Language Learning
412 30th Avenue Court
Greeley, CO 80631

Nancy Fairchild
Language Arts Teacher
Greater Albany School District
2310 S. Elm Street
Albany, OR 97321

Jack Fenlon
Coordinator of Program Evaluation
Cherry Creek School District
4700 S. Yosemite Street
Englewood, CO 80111

Eleanor Francke
Admin. Asst. Instruction
Lincoln Public Schools
Lincoln, NE 68505

R. Charles Garner
Program Director for Planning
and Evaluation
Region One Education Service Center
1900 West Schunior Street
Edinburg, TX 78539

Michael Glover
New Mexico Dept. of Education
Capitol Complex
Santa Fe, NM 87503

Sheri Haan
Christian Schools International
3350 E. Paris
Grand Rapids, MI 49508

C. Michael Hayes
Consultant
Westinghouse Data Score/Systems
770 Lucerne Drive
Sunnyvale, CA 94086

Al Hazelton
Supervisor of Research
Alaska Dept. of Education
Pouch F, State Office Bldg.
Juneau, AK 99811

James F. Hertzog
Basic Educ. Associate II
Pennsylvania Dept. of Educ.
111 Hanover Lane
Harrisburg, PA 17112

John Hills
Professor
Florida State University
1004 Lasswade Drive
Tallahassee, FL 32312

Michael Hiscox
Assessment & Measurement Program
NWREL
710 S.W. Second Avenue
Portland, OR 97204

David Holdzkom
Dissemination Specialist
Appalachia Educational Lab
P.O. Box 1348
Charleston, WV 25325

John Hollowell
Asst. Professor of English
English Department
University of Arizona
Tucson, AZ 85721

Susan Holt
Language Arts Consultant
Wyoming State Dept. of Educ.
Hathaway Building
Cheyenne, WY 82002

Ann Humes
SWREL
4665 Lampson Avenue
Los Alamitos, CA 90720

Carl I. Johnson
State Assessment Program
New Jersey Dept. of Educ.
225 West State Street
Trenton, NJ 08625

Laura Joplin
Colorado Women's College
Montview Blvd. & Quebec
Denver, CO 80220

John D. Kauffman
Vice President, Marketing
Scholastic Testing Services
480 Meyer Road
Bensenville, IL 60106

Stephen Koffler
Director, State Educ. Asmt.
New Jersey Dept. of Educ.
225 W. State Street
Trenton, NJ 08625

Anna Kogel
Educational Consultant
Jefferson City Schools
10960 Park Range Road
Littleton, CO 80123

Robert C. Korte
Supervisor of Assessment
Grand Rapids Public Schools
110 Ionia Avenue, N.W.
Grand Rapids, MI 49503

S. R. Lacher
Asst. Director
Information and Research
Dept. of Public Instruction
Bismarck, ND 58505

Joann B. Lipshires
Assoc. Director of Instruction
New Jersey Educ. Assn.
180 W. State Street
P.O. Box 1211
Trenton, NJ 08608

Robert Loveridge
1517 S. Theressa
St. Louis, MO 63104

George McCulley
English Education
Utah State University
UMC 28
Logan, UT 84321

Bernard McKenna
Senior Professional Associate
NEA
1201 16th Street, N.W.
Room 714
Washington, D.C. 20036

Lynn Mappus
Department of Education
Office of Res., Room 604
1429 Senate Street
Columbia, SC 29201

Jennifer Maryak
Evaluator
Ingham Intermediate School Dist.
Office of Planning & Evaluation
2630 W. Howell Road
Mason, MI 48854

Sandra K. Meacham
Assessment & Measurement Spec.
Central Valley School District
S. 123 Bowdish
Spokane, WA 99206

Sue Merkhofer
Secondary Language Arts Cons.
Florida Dept. of Education
Knott Building
Tallahassee, FL 32301

Thomas G. Montebell
Coord. of Assessment & Testing
West Virginia Dept. of Educ.
Bldg. 6, Room B057
1900 Washington Street
Charleston, WV 25305

Patricia D. Murphy
Institute for Teacher Education
North Dakota State University
Fargo, ND 58105

Donna L. Nola
Administrative Officer
Bureau of Accountability
Louisiana Dept. of Education
P.O. Box 44064
Baton Rouge, LA 70804

Betty Lou Pagel
Coord. Reading/Language Arts K-12
Laramie County School District #1
2810 House Avenue
Cheyenne, WY 82001

Elizabeth F. Penfield
Chairman, Freshman English
University of New Orleans
1204 County Road #170
Carsondale, CO 81623

Howard Potts
Asst. Admin., PRE Section
State Dept. of Education
2500 N. Lincoln Blvd.
Oklahoma City, OK 73105

Jerry M. Powers
1517 S. Theressa
St. Louis, MO 63104

Diane C. Proctor
Coordinator, Program Evaluation
Jefferson County Public Schools
1211 Quail Street
Lakewood, CO 80215

Linda Reed
Director
R & D Interpretation Service
CEMREL, Inc.
3120 59th Street
St. Louis, MO 63139

Agnes G. Rezler
Professor
University of Illinois
CED, 808 S. Wood
Chicago, IL 60612

Mel Riggs
English Consultant
Kansas State Dept. of Educ.
120 E. 10th
Topeka, KS 66612

Edward D. Roeber
Supervisor - MEAP
Michigan Dept. of Education
P.O. Box 30008
Lansing, MI 48909

Donald Rothman
Director
C. California Writing Project
Oakes College, UCSC
Santa Cruz, CA 95064

Henry W. Schultz
Coordinator
Measurement and Evaluation
Department of Education
#307-1181 Portage Avenue
Winnepeg, Canada
R3G 0T3

Cleona B. Shortridge
Coordinator
Oral/Written Comm. Project
CEMREL, Inc.
3120 59th Street
St. Louis, MO 63139

Herbert R. Steffens
Director, Planning & Evaluation
Nevada Dept. of Education
Capitol Complex
Carson City, NV 89710

Charlene G. Stogsdill
S/F Coordinator
Laramie County School Dist. #1
2810 House Avenue
Cheyenne, WY 82001

Ernest W. Strang
NTS Research Corporation
2634 Chapel Hill Blvd.
Durham, NC 27707

William Strong
English Education
Utah State University
UMC 28
Logan, UT 84321

Ellen Susnak
Assessment Specialist
Minnesota Dept. of Education
550 Cedar Street
St. Paul, MN 55101

James Thomas
Manager, Professional Scoring
Westinghouse DataScore Systems
P.O. Box 30
Iowa City, IA 52244

Philip F. Thornton
Coordinator-Evaluation
West Virginia Dept. of Educ.
Capitol Complex, Room 6B-057
Charleston, WV 25305

Danielle Tibbetts
Evaluator
Montana Office of Public Instruction
Room 106, The Capitol
Helena, MT 59601

Ronald M. Torgeson
Director-Information and Research
North Dakota Dept. of Public Inst.
State Capitol
Bismarck, ND 58505

Clark Trivett
Consultant
North Carolina Dept. of Public Inst.
217 W. Jones Street
Raleigh, NC 27611

Karen Underwood
Department of Education
Len B. Jordan Office Bldg.
650 West State Street
Boise, ID 83702

Ramon Veal
University of Georgia
Aderhold Hall
Athens, GA 30602

Michael Violano
Administrative Editor
Cambridge Book Company
888 Seventh Avenue
New York, NY 10019

Janet Wall
Coordinator
Research and Evaluation
DOD-Dependents Schools
Hoffman #1
2461 Eisenhower Avenue
Alexandria, VA 22331

Willard G. Warrington
Director
Undergraduate University Division
Michigan State University
Room 276, Bessey Hall
East Lansing, MI 48824

David J. Wright
Technical Services Coordinator
NAEP
1860 Lincoln Street
Suite 700
Denver, CO 80295

Stephanie Zimmerman
Westinghouse DataScore Systems
P.O. Box 30
Iowa City, IA 52244